What Do Others Think About This Comment? – Recommending Diverse and Relevant User Comments in Comment Sections

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Jan Lukas Steimann

geboren in

Witten

Düsseldorf, April 2025

aus dem Institut für Informatik
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Martin Mauve
2. Prof. Dr. Stefan Conrad

Tag der mündlichen Prüfung: 30.06.2025

This dissertation is dedicated to my wife, whose unwavering support accompanies me on my journey, to my parents, whose contributions made my university education possible in the first place, and to my friends and colleagues, whose advice and support were invaluable throughout my work on this thesis. I will be forever thankful.

Abstract

Comment sections of news websites serve as popular place for public debate, where users share their views. As the readers go through comments, they might find themselves wondering about various perspectives regarding a particular comment. For example, in a discussion about housing costs, a comment might suggest replacing old homes with efficient apartments, sparking the question of what other perspectives people might have. Yet, uncovering such comments involves combing through related articles of several news outlets, which would take a lot of effort.

This dissertation presents a novel approach for recommending relevant and diverse comments that complement the comment the user is currently interested in. This research seeks to aid online discussions by offering users perspectives they might not typically encounter, thus fostering a more informed opinion. As this comment-centric approach is unique so far, this work investigates how to develop such a system and examines the challenges that are encountered. To ensure the practical value of this solution, we keep a strong focus on usability during this thesis.

This task of finding relevant and diverse recommendations presents a complex problem which is broken down into more manageable sub-tasks during this thesis. First, a method is discussed to identify recommendations that align thematically using semantic similarity. Concurrently, the retrieval process performance was evaluated to ensure the approach remained application-focused.

Subsequently, this thesis focuses on the application-oriented facet and presents an open-source software framework designed to swiftly build prototypes for the comment recommendation systems. These systems allow for an evaluation of the models in both laboratory and real-world environments.

Next, the thesis aims to identify relevant comments. Although initial research helps detect topically related comments, it may not ensure their relevance to the discussion. Thus, a user study was conducted to explore context-independent factors such as position or justification that may affect the relevance of a comment. This user-study helps to build an understanding of how these elements influence comment relevance.

In conclusion, these findings of previous studies are integrated and expanded to construct the final recommendation model. This marks the first instance of delivering relevant and varied recommendations for a specific comment, while also helping to develop an understanding of open-problems and possible solutions for future recommendation systems.

Zusammenfassung

Die Kommentarspalten von Nachrichtenseiten bieten Nutzern die Möglichkeit sich über ihre Ansichten und Meinungen zu einem Artikel oder Thema auszutauschen. Beim Lesen eines Kommentars fragt sich ein Leser vielleicht, wie andere darüber denken. Zum Beispiel könnte bei einer Diskussion über den Wohnungsmarkt ein Nutzer vorschlagen, alte Häuser durch neue, energieeffizientere zu ersetzen. Das könnte die Frage beim Leser aufwerfen, was für andere Meinungen dazu existieren. Es ist jedoch äußerst aufwendig solche Kommentare zu finden, da der Leser dazu zahlreiche Artikel verschiedener Nachrichtenagenturen durchsuchen müsste.

Diese Dissertation stellt einen neuartigen Ansatz für die Empfehlung von Kommentaren vor, der darauf abzielt, relevante und vielfältige Kommentare zu identifizieren, basierend auf dem Kommentar für den sich der Nutzer gerade interessiert. Dabei ist das Ziel Meinungen und Perspektiven aufzuzeigen, welchen der Leser normalerweise nicht begegnen würde. Auf diese Weise soll der Nutzer sich eine ausgewogenere Meinungen bilden können. Da es sich hierbei um einen neuen Ansatz handelt, untersucht die Arbeit vorwiegend, wie ein solches System anwendungsorientiert entwickelt werden kann und welche Herausforderungen und Lösungen dabei auftauchen.

Um diese Herausforderung zu meistern, wird die komplexe Aufgabe des Empfehlens von Kommentaren zunächst in kleinere Teilschritte zerlegt. Zuerst befasst sich diese Arbeit mit der Schwierigkeit, relevante Kommentare basierend auf semantischer Ähnlichkeit zu ermitteln, die thematisch zu dem Kommentar passen, für den sich der Nutzer interessiert.

Anschließend wird der anwendungsorientierte Aspekt der Arbeit genauer betrachtet. Hierfür wird ein Framework entwickelt, welches darauf ausgerichtet ist, schnell Prototypen für Empfehlungsmodelle zu bauen. Diese lassen sich dann sowohl in einer Laborumgebungen als auch im realen Kontexten evaluieren.

Als nächstes befasst sich die Dissertation mit der Identifizierung relevanter Kommentare. Während die erste Forschungsarbeit hilft, thematisch verwandte Kommentare zu erkennen, kann sie nicht gewährleisten, dass der Kommentare einen relevanten Beitrag für die Diskussion liefert. Daher wird eine Benutzerbefragung durchgeführt, um zu untersuchen, welchen Einfluss kontextunabhängige Faktoren wie Position, Begründung und Quellen auf die Relevanz eines Kommentars haben.

Abschließend werden diese Erkenntnisse aus früheren Studien in einem Modell gebündelt, welches als erstes automatisiert relevante und vielfältige Kommentaren für einen gegebenen Kommentar identifizieren kann. Gleichzeitig konnte dadurch ein Verständnis für offene Probleme und mögliche Lösungen für zukünftige Empfehlungssystem entwickelt werden.

Acknowledgements

My heartfelt appreciation goes to many who made this dissertation possible. Firstly, I thank Martin Mauve for allowing me to work and research under his guidance. Special thanks to colleagues Markus Brenneis and Marc Feger for their help, advice, and for listening to my random ideas. My gratitude extends to Thomas Spitzlei for helping me in getting my applications online. I am grateful to other colleagues who contributed through discussions, comments, and social interactions, despite the constraints of Covid-19. Thank you all: Björn Ebbinghaus, Simon Weber, Katharina Boland, Jens Bendisposto, Janine Golov, Philip Körner, Raphael Bialon, Sabine Freese, Lisa Lorenz, and many others I might have overlooked. Undoubtedly, my deepest gratitude goes to my parents and my wife for their unwavering support over the years.

Contents

Chapter 1

Introduction and Motivation

The rise of the World Wide Web has significantly transformed the manner in which individuals engage in discussions. Merely a few decades earlier, opportunities for discourse about current events were largely confined to interactions with family, friends, or casual conversation at local pubs or cafes. Presently, however, individuals can articulate their perspectives on any news story by commenting below the respective article or post online, as well as respond to the opinions shared by others on the subject. There have been, and continue to be, aspirations that this capability might foster a more enlightened society due to the ease with which individuals can exchange views and nurture a more profound comprehension of each other's perspectives. Despite these hopes, many continue to behave as they would within traditional offline environments, predominantly frequenting websites affiliated with their favored news sources (Flaxman et al., 2016). Consequently, a significant number of individuals are primarily exposed to opinions that align with their own political leanings. Furthermore, even if one is inclined to explore diverse viewpoints concerning the comment they are currently examining, identifying these alternative insights poses a challenge. Where might they seek such diverse perspectives? Which news organizations should they consider? Moreover, amidst the flood of comments, how might they pinpoint responses that specifically address the issue raised in the comment they are reading?

In this thesis, we thoroughly examine the area of comment recommendation to support the ongoing research, discussed in Section 2.3. Our aim is to recommend user comments with different points of view for the topic of the given comment the user is currently interested in. To the best of our knowledge, we are the first to address this subject from a comment-centric, detailed perspective, contrasting with previous studies which, e.g., are based on personal interests of the user (Agarwal et al., 2011; Zhou et al., 2015a) or try to identify the most important or constructive comments in the overall discussion (Kobayashi et al., 2021; Mahajan et al., 2012). To the best of our knowledge, we are the first to propose recommending comments based on another comment the user is currently interested in. Through this thesis, we aim to introduce this novel recommendation task and establish a benchmark, assessing the existing research's capabilities and pinpointing challenges and potential solutions for future endeavors.

News outlet comment sections typically present comments in an ordered list, prioritizing the most replied to or recent posts; a practice largely unchanged over recent years. This leaves many challenges unaddressed, such as helping users navigate the overwhelming volume of comments.

Despite extensive research in this area, detailed in Section 2.3, the paper in Section 4 will explain that much of the research emphasizes developing sophisticated models rather than focusing on application-focused approaches. This might contribute to the lack of innovation in how to apply existing research in the comment sections.

This thesis takes a practical-focused approach to address this problem. To build a recommendation system which allows for a real-world use of our recommendation model, several challenges must be overcome. Firstly, how can we quickly locate comments that align with the themes the comment is addressing? How can we maintain an application-centered approach during the system's development? How do we select comments that offer valuable contributions to discussions? Lastly, how can we identify diverse perspectives to present users with various points of view?

1.1 Research Questions

In this section, we delve deeper into the previously posed questions. We want to identify suitable recommendations that not only offer relevant positions but also bring fresh perspectives by presenting diverse viewpoints while keeping the practical-orientated aspects in mind.

How can we find thematically fitting comments in a reasonable amount of time? When considering how to generate appropriate recommendations, our initial consideration is to identify comments that align thematically with the comment the user is interested in. However, due to the practical nature of the application, this is just a partial aspect of the challenge. It is equally important to ensure that these recommendations are obtained in a reasonable amount of time. Thematic alignment and performance are the fundamental requirements that our recommendation model must satisfy. Even the most sophisticated and persuasive arguments in the world are meaningless if they do not resonate thematically. Likewise, recommendations become useless if they cannot be retrieved and presented to the user promptly.

How can we develop recommendation system prototypes with the practical focus in mind? As previously mentioned, this thesis is significantly shaped by the usability and practical application of our recommendation system. Consequently, we must devise a method to ensure that our proposed recommendation models can be assessed both in realistic environments and laboratory conditions. This approach is necessary to confirm that our model can offer its recommendations within a reasonable time-frame.

How can we identify relevant user comments? After determining how to find comments that are thematically suitable to the user's comment, it is essential to assess if a possible recommendation also presents a robust stance and reasoning. Recommending user comments that align thematically yet lack a substantial position or rationale in the discussion would be futile, as they offer no new insights to the user.

How can we combine these insights with identifying diverse perspectives to present suitable recommendations to the user? Upon addressing all the previously posed questions, the challenge lies in integrating these insights to construct a recommendation model. This model should not only suggest comments that are thematically and argumentatively pertinent but also ensure that the recommended comments encompass a variety of perspectives. This diversity enables users to consider multiple viewpoints and gain a comprehensive understanding of the subject in the comments they are interested in.

1.2 Contributions and Outline of This Thesis

In the next chapter, we explain our definition of a relevant user comment and review the existing literature and describe how our study integrates within the current research landscape.

In Chapter 3, we introduce a straightforward yet effective recommendation model aimed at tackling the first research question regarding the rapid detection of thematically relevant comments. This model employs semantic similarity with precomputed vector embeddings of the comments within our database, calculated during the import phase to align with our practical focus. During this process, we also incorporate the articles related to the comments and utilize the article's keywords to find articles pertinent to the user comment. Consequently, we limit our focus to comments associated with these articles to improve retrieval speed while considering the thesis's practical application objectives. Although this model offers reliable recommendations, it is insufficient to comprehensively fulfill the recommendation task we anticipate. However, it marks an initial step towards developing a universally applicable recommendation system. To succeed in making meaningful recommendations, it is essential not only to achieve semantic alignment but also to incorporate robust argumentation and personal perspectives, along with other attributes, in order to effectively meet the needs of our recommendation task. Simultaneously, these recommendations need to be assessed in real-world scenarios to maintain a practical emphasis.

Therefore in Chapter 4, we proceed to explore the second research question. We present an adaptable, open-source framework designed for the efficient creation of prototypes aimed at recommending comments. This framework is composed of a comprehensive ecosystem of components that together offer the necessary technical infrastructure to develop prototypes with high efficacy, both in controlled laboratory environments and in practical applications in real-life contexts. Through this paper (Steimann and Mauve, 2024), we aim to underscore our research's practical orientation and applicability. This inquiry lays the foundation for developing the recommendation model that we elaborate upon in Chapter 6.

Subsequently, in Chapter 5, we explored the impact of structural features such as position, justification, and personal narratives on the perceived relevance of user comments (Steimann et al., 2024). Our objective was to gain insights into how these structural components affect the comment's perceived relevance and to determine if their presence or absence alters the relevance of user comments. To achieve this, we conducted a user study in which participants were presented with identical comments, differing only by one structural element, and asked to judge which comment they found more or most relevant. This understanding is crucial in Chapter 6, where we develop a score to assess a comment's relevance independently of its context, which can be precomputed during database import.

Chapter 6 builds on our previous research, addressing the last research question by introducing our recommendation model, which integrates the insights of all previous studies. This model generates recommendations aligned with user comments, offering relevant and diverse suggestions. Initial evaluations exhibit promising results, with the model successfully uncovering pertinent and varied opinions. Nonetheless, these findings are preliminary due to the evaluation's limited scale. Several challenges remain, particularly the restriction to a limited number of topics and the inability to distinguish between factual and false information, which require future attention. We also need to find ways to better evaluate future models.

Finally, in Chapter 7, we consider the foundation established by this thesis for the novel recommendation task discussed earlier in this chapter. We evaluate the recommendation model that can be developed using contemporary research methods and identify challenges and solutions for future endeavors. Consequently, we pause to examine the successes and difficulties encountered in this thesis, while outlining future challenges and potential remedies.

Chapter 8 presents a concise overview of our study and the key observations that conclude the dissertation. Additionally, we extend a summary of our perspectives on future work discussed in Chapter 7.

Chapter 2

Preliminaries and Related Work

The field of recommending online news content has been studied extensively. In this chapter, we begin by explaining the definitions of a relevant comment and a diverse and well-rounded set of viewpoints, which are based on previous research. Following these definitions, we will then examine the current research in news content recommendation and explain how the work presented in this thesis integrates into this context.

2.1 Definition of the Term “Relevant Comment”

To recommend relevant and varied comments related to the one that a user is focused on, we need to discern which comments hold relevance and which do not. Hence, we outline the criteria for what constitutes a relevant comment below. This definition is drawn from previous research on the subject and is investigated in more detail in Chapter 5.

Beforehand, it is vital to clarify our goals in recommending comments. We aim to aid users who engage in public discussions within online news agency communities. Nevertheless, as highlighted in Chapter 1, a significant number of users interact with digital media as they do with conventional platforms, mainly frequenting the websites of their favored news organizations (Flaxman et al., 2016), thereby predominantly encountering perspectives from the same community. Consequently, our objective is to suggest insightful comments that users might not typically encounter, as such viewpoints are more common in different communities, fostering a more well-rounded opinion. Therefore, the comments we provide should encompass diverse viewpoints, not merely well-structured arguments, on the subject. In this context, comments that offer alternative perspectives backed by solid reasoning, including both factual information and personal experiences, are more valuable as they enable users to appreciate a wider range of opinions.

2.1.1 Relevant Comments

The following section delves into the definition of the characteristics that influence the significance of a user comment. It further details which characteristics are deemed necessary for a comment to be relevant and also how this relevance is impacted by the inclusion or exclusion of certain structural elements. For our purposes, we utilize the characterization of a relevant user comment as detailed in Rowe (2015):

First, for a comment to be deemed relevant at all, it needs to align with the core subject. A comment straying from the topic of interest to the user is unlikely to be considered pertinent. Rowe (2015) classifies topic relevance into two main categories. The first category is the structuring topic, which includes comments that are explicitly centered on the principal topic of the discourse. The second category is the interactional topic, involving comments that, while addressing a secondary but related issue, still maintain relevance within the broader context of the dialogue. For instance, in discussions on immigration, mentions of economic concerns frequently emerge, making such comments pertinent as well.

Rowe (2015) additionally considered *opinion direction* in their study, which indicates the ideological direction either as *conservative* or *liberal.* This allowed the authors to assess the diversity of comments in their study. We also include diversity of opinion in this thesis. However, we follow a different interpretation, in which we expect a spectrum of opinion ranging from *conservative* to *liberal.* Rowe (2015) follow a more binary interpretation where the comment is either *conservative* or *liberal.*

A comment being *on topic* is just a prerequisite. For comments to be relevant, they also need to present their position on a given issue or policy with a justification, allowing others to evaluate the argument. A simple statement without additional interpretation or opinions does not provide new perspectives to the reader. For example, the following comment is usually not considered relevant: *The author of this article is wrong and has no idea about the topic.*

Although the characteristics detailed in the preceding paragraph establish a baseline for a meaningful comment, Rowe (2015) noted further traits which are considered to influence perceived relevance. For instance, when comments integrate and reference supplementary sources, they enable readers to assess the justification's reliability and accuracy, thereby boosting the comment's importance and relevance. Furthermore, personal anecdotes within comments can help bridge communication gaps resulting from complex subjects or insufficient comprehension by presenting the issue in a more relatable manner. These narratives effectively break down such barriers. Additionally, comments that suggest alternative solutions to the problem are deemed more pertinent than those that do not, as they offer a novel viewpoint on the issue. For instance, discussing the pros and cons of renovating old houses versus purchasing a new one can be more compelling if it includes personal experiences rather than solely focusing on cost-effectiveness. Other structural elements considered by Rowe (2015) are *interaction*, where participants consider and respond to others' perspectives, or *question*, where participants seek to understand others' viewpoints through inquiry.

For our purpose, not all of these structural elements are useful. As explained in the beginning of this chapter, we aim to recommend comments from different communities. Therefore, specific elements like *question* or *interaction* are intentionally omitted by us as these characteristics carry more weight in ongoing discussions since they focus on the interaction among participants and are less critical for our purposes. Comments that are considered significant in our context should remain relevant to the subject, regardless of the context of the discussion. Our goal is to draw comments from varied discussions and media outlets to showcase diverse perspectives on the issue.

2.1.2 Populist Messages

There is another aspect of comments which is very important in online discussion and which is difficult to classify: *populist messages.*

Populist rhetoric can often be identified by several key themes (Sina Blassnig and Esser, 2019). Rather than concentrating on the issue itself, they emphasize *people centrism*, aiming to connect with audiences by highlighting their virtues and accomplishments, portraying the populace as unified and dominant. Concurrently, they marginalize certain groups, attributing problems to them, frequently coupling these claims with anti-elitist sentiments.

The challenges posed by this category of comments arise from their dual nature: they can disseminate populist messages that may disrupt discussions, yet they might also meet our relevance criteria as defined in Section 2.1.1. This dilemma is inherent in all comment recommendation systems. The system must recommend engaging comments to assist users in sifting through the overwhelming volume of postings without being exploited for propagating hateful content. Thus, these comment types must be considered in the development of a systems which faces the public. Identifying these messages is a complex research endeavor requiring careful evaluation, as distinguishing between populist rhetoric and regular user opinions can be difficult. Consequently, we have excluded this issue for now but propose initial strategies to address these comments in Section 7.2.

2.2 Diverse Selection of Comments

The purpose of this thesis is to broaden users' horizons by delivering pertinent commentary that introduces fresh insights and unexplored viewpoints. To achieve this goal, our selection process needs to include a range of ideological perspectives related to the topic of the comment. This means that our selection should encompass comments that align with the comments' position while also offering new arguments or insights concerning the topic. The ideal selection of comments would include a spectrum of opinions for the given comment, ranging from opposition to agreement and everything in between. Our aim is not to present a dichotomy of opposing perspectives debating what is correct or incorrect. Instead, we seek a collection of diverse viewpoints, supported by individual insights and sources, to promote a deeper comprehension of alternative opinions on the topic, which people might not typically encounter. For example, the user's comment focuses on installing solar power panels on their house. The optimal compilation of comments should feature those supporting the viewpoint, highlighting both environmental and financial advantages they had with the installation on their house, as well as comments adopting a more neutral stance that delve into the nuances of cost-benefit elements.

In addition, it should include opinions opposing solar power panels, citing concerns such as unreliability and durability issues.

2.3 Related Work

Various research papers deal with recommendations for the content of news agencies. Probably, the most well known intended purpose is the recommendation of news articles which are of interest to the user (Li et al., 2010; Liu et al., 2024). However, this is only one possibility where a recommendation system might help the user deal with the flood of content online. If we take into account the potential for users to engage in public discourse via discussions in comment sections, we can identify various applications for recommendation systems, such as identifying significant contributions or tailoring recommendations to match the user's personal interests. Nevertheless this also brings challenges that need to be dealt with to allow for a constructive debate.

Similar to the recommendation of news article where people have trouble to find the content they might be interested in, they also have trouble finding the discussions they might have an interest commenting. Therefore one research approach was to develop personalized recommendation systems that recommend news articles with discussion that the user probably will write comments in (Risch et al., 2020; Shmueli et al., 2012). For example, Risch et al. (2020) developed a neural network which learns embeddings for users and comments to compute a probability if a user posts a comment in the given discussion.

Yet, on its own, this is insufficient for tackling the issues associated with online commenting like navigating the flood of comments. Therefore, another strategy is to directly address the overwhelming volume of comments and to assist users in navigating through this effectively.

For instance, Agarwal et al. (2011) tries to sort the comments in a discussion for the user and developed a model which ranks the comments associated with a specific article based the probability that the user might like or dislike the comment. For this, they compute the similarity of viewpoints between the user and the comment and the user and the author of the comment. Another approach was taken by Zhou et al. (2015a) where the authors developed a multidimensional classification and personalization system that recommends comments based on text features and personal interest of the user.

An alternative method to help users navigate the flood of comments, used by many publications, is to identify the comments that contribute constructively in the discussions, independent of the interests of the user. There are essentially two types of approaches for this.

A potential solution involves aiding moderators employed by news agencies. These moderators are tasked with ensuring respectful discourse and recognizing valuable contributions. Despite their role, they encounter the same challenge as regular users—overwhelmed by the sheer volume of content needing review. Consequently, certain publications (Park et al., 2016;

Waterschoot and Bosch, 2024) strive to assist moderators by pre-processing comments and offering supplementary information to streamline their tasks.

Park et al. (2016), for example, combines a trained classifier which ranks the comments based on various features like readability, relevance for the article, etc. with a user-interface to help the moderators to decide if a comment should be approved, rejected or even highlighted as a high-quality contribution. These decisions are then used as feedback for the classifier to adapt it more to the news agency's expectations for good comments.

The other approach is to present the results directly to the user by ranking the comments based on how constructive or relevant they are for the discussion (Kobayashi et al., 2021; Kolhatkar and Taboada, 2017; Kolhatkar et al., 2020; Mahajan et al., 2012; Uribe et al., 2020).

A study by Kolhatkar and Taboada (2017) compiles a dataset comprising NyTimes Picks, which are comments highlighted by NyTimes moderators for their exceptional quality or originality in perspective. This dataset is then complemented with non-constructive comments from the Yahoo News Annotated Corpus serving as negative samples. Subsequently, the authors assess several classification models utilizing SVMs and bidirectional LSTMs implemented with diverse features. Their most successful SVM model achieves an F1 score of 0.84, while the top-performing LSTM model attains an F1 score of 0.81. The researchers emphasize these results as preliminary, illustrating the potential for automating the identification of constructive comments.

Mahajan et al. (2012) uses a logistic regression model on word and topic features to predict the average rating for a comment and proposes to sort the comments accordingly to this value to provide the user with the top-N comments. However, the authors note a very important aspect of comment recommendation, ranking comments only based on their constructiveness carries the risk of recommending only very similar or redundant comments.

However, diversity of opinions is not a *nice to have* feature for comment recommendation, but a substantial aspect that needs to be considered. As Joris et al. (2020) has already noted, the importance of creating a diverse news and opinion offering has been recognized by lawmakers in several policy documents and the possibility for citizens to receipt these is an important prerequisite for a democratic society. Zerback and Schneiders (2024) further demonstrated that news audiences highly value the inclusion of multiple viewpoints in news content. They perceive sources offering a broader range of perspectives as possessing greater credibility.

Many research methodologies acknowledge the importance of diversity and strive to integrate it into their models. For instance, Mullick et al. (2019) approaches this by hypothesizing that the comment section beneath a news article consists not of a singular discourse but numerous smaller conversations, each relating to distinct segments and themes within the article. To address these smaller conversations, they have designed a deep neural network that links newly published comments to particular sections of the article. In conjunction, they provide a user interface that displays only those comments relevant to the selected section. This association enables users to gain a more thorough understanding of the diverse discussions occurring within the comment section, effectively illustrating varied aspects of the topic at hand. However, this model provides and interesting approach for comment recommendation. However, it does not address all challenges. For instance, as the authors acknowledge, a comment may pertain to multiple sections. Additionally, even when comments are filtered to a specific section, that section might still contain an overwhelming number of comments, potentially offering limited diversity of opinion. Consequently, there remains a need to implement a recommendation system capable of ranking comments and ensuring a variety of perspectives.

Other research approaches try to connect the comments from different sources to provide the users with a more diverse set of opinions.

The work by Kim et al. (2021) seeks to bring awareness to diverse viewpoints on a subject by proposing the use of *Hagendas*, which merge hashtags with agendas. This concept allows users to engage in related discussions across multiple platforms such as comment sections and social media. The authors aim to link these discussions by connecting new comments to keywords that are automatically extracted from the headline and initial sentence of an article. In this way, users are enabled to explore all comments linked to a particular hagenda, thereby gaining insight into all viewpoints associated with the chosen keyword. In their study, Kim et al. (2021) propose a interesting method to link comments from a variety of sources. Despite this, they encounter comparable challenges to those identified by Mullick et al. (2019), particularly related to the volume of comments that could pertain to a particular agenda. Users are limited to exploring overarching topics such as *US* or *espionage*, which are likely to be connected with an extensive number of comments. Consequently, a ranking mechanism for these comments is necessary to effectively manage the overwhelming influx and to ensure a diversity of perspectives is maintained.

An intriguing approach detailed in Risch et al. (2021) parallels that in Kim et al. (2021) by presenting multiple perspectives through the organization of comments from various news outlets into a graphical format, which is rendered visible to users via a specially designed interface. This process requires the examination of every new comment, treating each sentence as an individual node within the graph. Connections among these nodes are formed based on thematic similarities and other attributes of the comments. These attributes aid in deciding the weights and labels for the edges, which are then used to merge and cluster nodes, offering a detailed and comprehensive summary.

The studies by Risch et al. (2020) and Kim et al. (2021) primarily concentrate on linking and exploring topics and comments across diverse media outlets using tailored user interfaces. These interfaces are designed to provide users with a comprehensive understanding of various topics, thus contributing significantly to public discourse.

Alternative research approaches, as discussed by Chen et al. (2019a,b), have addressed diversity concerns by providing multiple perspectives on a specific claim, supported by evidence. Chen et al. (2019b) introduced the task of *substantiated perspective discovery*, which includes sub-tasks such as *Perspective Extraction, Perspective Stance Classification, Perspective Equivalence, and Extraction of Supporting Evidences*, and offered a dataset containing claims, perspectives, and evidence. Later, Chen et al. (2019a) developed a web interface for the dataset from Chen et al.

(2019b), which allows users to search a claim and obtain supporting and opposing perspectives with evidence.

As we have seen, research provides a variety of different approaches to aid users navigate the online news landscape. Users can explore news articles and comments, query various claims and obtain evidence-backed perspectives, and identify the most constructive comments in a discussion.

Our contribution to this subject is reflected in our effort to address the question in the title of this thesis: *What do others think about this comment?* Prior work, such as Chen et al. (2019a,b), employs a method akin to ours by offering varied perspectives on a given claim. However, our focus and methodology differ. Unlike Chen et al. (2019b), which sources perspectives from debate websites about a claim, we concentrate on user comments found in news article comment sections which are often less organized. While debate arguments and comments share similarities, they often differ in style—formal debates often feature more structured arguments. Though similar structured comments can be found in comment sections, users there usually focus more on social interaction (Diakopoulos and Naaman, 2011) like personal viewpoints related to the article's subject. We aim to capture these detailed arguments as well as show readers public opinions on the topic backed by personal views, while Chen et al. (2019b) emphasizes more on extracting well-supported arguments for specific claims.

Kim et al. (2021) and Risch et al. (2021) which try to connect comments from different sources and by this incorporating different points of views, also adopt a comparable method to ours. However in contrast to us, they focus on the overview and exploration aspects of comment recommendation. Our aim is to offer users an in-depth perspective in the discussion with different viewpoints for the given comment at hand and therefore provide a complement to this exploration research. Our research seeks to counter echo chambers and filter bubbles, encouraging deeper empathy in dialogues by offering a range of perspectives and personal insights on comments that engage user interest.

Chapter 3

An Efficient Method for Identifying Thematically Relevant Comments

This chapter provides a summary of the contributions of our paper Steimann et al. (2022) to identify thematically relevant comments in a timely manner:

Jan Steimann, Marc Feger and Martin Mauve.

"Inspiring Heterogeneous Perspectives in News Media Comment Sections"

In: *Proceedings of the 24^{th} International Conference on Human-Computer Interaction*, in the thematic area *Human Interface and the Management of Information: Visual and Information Design*, pages 118 – 131, Springer International Publishing
Acceptance Rate: ∼57%

We introduce the novel task of suggesting comments that are relevant to the comment the user is currently interested in, along with a preliminary model that employs semantic similarity to identify suitable comments. This paper marks the beginning of our research to tackle this issue. For this reason, we find an early definition for relevance in this paper which differs from our definition in section 2.1.1.

3.1 Summary

This paper introduces a novel task in comment recommendation along with a straightforward model to address a challenge encountered in pursuing this task. The objective is to identify comments thematically related to a comment currently of interest to a user, enabling them to gain insights into others' perspectives on the subject.

The introduction explores the challenge of limited diversity within comment sections and illustrates the motivation behind our novel recommendation task. We propose to tackle this issue by redirecting the focus of comment recommendations from general user preferences to the specific comments and topics that intrigue them. To achieve this, we introduce the idea of an ecosystem with multiple components that collect comments from diverse news sources and communities and offer recommendations centered on the user's current comment of interest.

Our aim is to deliver a set of relevant yet diverse perspectives to the user. This paper marks the beginning of this ecosystem's development by presenting a model to efficiently identify comments that align with a given user comment.

In the related work section, we dive into other research approaches for comment recommendation. There we detail the distinctions between our approach and earlier studies and how our method intends to complement current approaches by finding thematically relevant comments. With this, our goal is to help users during the opinion-forming process.

Section 3 of the paper outlines the process through which the model efficiently pinpoints recommendations by employing a two-step method that utilizes sentence embeddings and cosine similarity. In the following section of the article, we describe the evaluation process of our model. In this section, we provide an in-depth account of our trials with different embedding models to identify the optimal performer for our specific objective. As we wrap up the paper, we delve into the evaluation outcomes and potential directions for future research.

3.2 Personal Contribution

The initial idea of recommending comments based on other comments was introduced by Martin Mauve. The thesis author, Jan Steimann, further advanced this idea, creating the idea of an ecosystem of components to suggest comments from a variety of communities. He engineered the recommendation model, devised the two-step method for making timely recommendations, designed and carried out the evaluation, and authored the full paper. Marc Feger provided the software utilized for the evaluation and recommended using Kripendorff's Alpha to measure inter-annotator agreement.

3.3 Importance and Impact on this Thesis

This paper initiates our thesis' exploration of the research questions described in Section 1.1, by recommending comments for a specific user comment via vector embeddings and semantic similarity. To the best of our knowledge, this is the first effort to offer comment recommendations within the scope of news article discussions based on other comments. It introduces our preliminary model, designed to quickly identify pertinent comments thematically related to the given user comment. This model is the base on which we build in Chapter 6 where we present our final recommendation model, which synthesizes all other previous research papers in this thesis.

Inspiring Heterogeneous Perspectives in News Media Comment Sections

Jan Steimann( ), Marc Feger, and Martin Mauve

Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany
jan.steimann@hhu.de

Abstract. Discussions in the comment sections of news articles are often characterized by highly one-sided arguments. One reason for this is that users primarily consume information that fit into their worldview and consequently argue from their echo chamber when responding to a comment. In this paper we introduce a new approach to comment recommendation. To this end, we present a model that makes recommendations based on a specific comment the user is interested in. These recommendations provide the user with alternative related comments and thereby broaden the perspective of the user before reacting to the specific comment that was originally selected. This is in contrast to previous work that tries to recommend comments based on the interests and previous behavior of the user and therefore fueling the filter bubble by providing information that fit into the worldview of the user.

Keywords: Artificial intelligence and IoT · Evaluation methods and techniques · Information bubble and echo chamber

1 Introduction

In the comment section of news media, people exchange arguments and opinions on the content of the article they have read. These comment sections are often characterized by highly one-sided arguments due to the reason that news agencies often attract reader with a political orientation that coincide to their own [4]. Therefore, we have a rather homogeneous argumentation in the comment section and people tend to behave differently as in a heterogeneous environment [1,2,8]. We believe that comment sections would be much more interesting and beneficial to the user and the society if they would include more diversity.

Our approach to solve this problem is to provide a heterogeneous but well balanced selection of comments from different discussions of other articles and news agencies based on a specific comment the user has selected. This way, we do not focus on the user but rather on the comment for our recommendation. To provide this new perspectives for the discussion, we are developing an ecosystem of components to extract relevant comments from a database for the comment the user is interested in real-time. For this, we create a database that is updated on a regular basis with articles and comments from various news agencies of

S. Yamamoto and H. Mori (Eds.): HCII 2022, LNCS 13305, pp. 118–131, 2022.
https://doi.org/10.1007/978-3-031-06424-1_10

the political spectrum so that the system can provide the user with the latest comments.

In this paper, we propose a solution for one problem of this ecosystem. How to extract and present a relevant selection of comments given a individual comment selected by the user from a large dataset of comments in a reasonable time. Therefore, we present our model that makes recommendations based on the selected comment, i.e. a comment-centric-comment recommendation.

In the following chapter, we take look at related work in the field of comment recommendation. The third chapter will explain our model. After this, we discuss the evaluation of the model and its limitations.

2 Related Work

Previous work in the field of comment recommendation, such as [7,12,13], present a selection of the most interesting comments for the user based on the previous behavior and interests of the user. This way, it is easier for the user to navigate the flood of comments that are posted every day and to find discussions where they can share their expertise.

[7] suggested to promote the general quality of online debates in the comment sections by filtering low quality comments and identifying and highlighting high quality contributions for the user. They try to achieve this by developing a comment moderation system with a flexible interface that interactively identifies high quality comments for the moderator. For this, the comments are scored by several quality criteria like *Article Relevance, Conversational Relevance, Readability etc.*. Afterwards, the comments are ranked by these scores and presented to the moderator with additional meta information, who decides if the comment is accepted, rejected or highlighted as a high quality contribution. Different from our work, the authors focus on a tool to assist the moderator of the comment section of a news agency to find high quality comments in a discussion. Our model tries to suggest highly relevant comments based on the comment the user is interested in. To do this, we search various discussions that might have happened a while ago but still contain relevant perspectives and ideas that differ from the perspective of the comment the user is interested. By this, we provide the user the possibility to build upon those ideas formulated by other users that have a different view on the topic and by this write a much more sophisticated response.

[13] presented an approach for a personalized news comment recommendation system to navigate the flood of comments. They designed a three part system to find the most relevant comments for the user. In the first phase, they filter out comments which contain much repeated or unrelated information by searching for comments that contain joking words, advertisements, and repeating words. Following, the system classifies the comments in *insightful view* or *informal comment* by various dimensions like similarity between the comment and the article, the length of the comment, or the attitude. In the final step, the system provides personalized recommendations based on the record of the user which comment

he or she wrote or liked. In our work, however, we focus on providing comments to the user based on the comment they are interested in. We believe if the user is presented comments only based on their record, it will most likely enforce the filter bubble and echo chamber which in return leads to a one-sided argumentation. The user can formulate a much more thoughtful response if they consider different opinions for the topic of the discussion.

[12] discussed different design goals a personalized comment section system should fulfill. The author considers three main points such a system needs to address to avoid user concerns like filter bubbles or missing the bigger picture. First, any comment proposed to a user needs to be relevant for the user. Second, to avoid being trapped in the filter bubble, the proposed comments need to consider different point of views. At last, the system should also help the user to find comments where they can share their expertise and also point to comments that could deepen their knowledge about the topic of the article. With our work, we try to complement prior work in the field of comment recommendation which proposes comments to the user based on their interest, like [13]. Such work fulfills the first design goal to find relevant comments for the user and to find comments where they can share their expertise. Our model addresses the rest of the design goals by offering different point of views to combat the filter bubble and allow the user to gain new insights through other user comments.

3 Model to Find Relevant Comments

In the following section, we introduce our model to find relevant comments based on the comment **C** of the current discussion the user would like to argue about. In our future work, this model will be part of a greater ecosystem that should support the user while debating in the comment sections. In this ecosystem the model will extract relevant comments for **C** from a database that will be updated on a regular basis with comments from discussions from various news agencies.

In order to suggest comments to the user based on the comment they are interested in, and to do this sufficiently quickly, so that they can use the suggestions to formulate their response, it is critical that the model can make the suggestions in an a short time.

For this reason, our model[1] consists of a two-step approach. In the first, step the model makes a preselection to reduce the amount of possible comments to a manageable amount and therefore generates a candidate set of possible comments from which we extract the best suggestions for the user in the second step, like [3] for argument search.

To generate the candidate set, we first consider the article **A** for the discussion the user is currently participating in. We use the article to find other discussions that deal with the same topic. For example, arguments and opinions about climate change under a article about renewable energies could be useful for an article about the harmful effects of a coal-fired powered plants. Likewise,

[1] https://github.com/hhucn/Inspiring-Heterogeneous-Perspectives-in-News-Media-Comment-Sections.

articles that seek to explain a need for coal power plants may be relevant, as they are likely to have different opinions and perspectives that will add value to a discussion about renewable energies. Of course, not every comment from these discussions will be useful for our current discussion but it will help to reduce the amount of possible candidates to a more manageable amount. To create the candidate set, we use the keywords of the article **A** under which the comment **C** appeared for which we want to generate suggestions. The keywords summarize the content of the article and therefore, we can use them to finde articles with a similar content. First, we embed the keywords of all articles in a vectorspace. For this, we calculate the average vector for every keyword and use a k-nearest neighbor search to finde the most similar articles to our article **A** (Fig. 1 left). Afterwards, we store all comments that have appeared under these k articles in the candidateset for the second step where we extract the n most suitable suggestions.

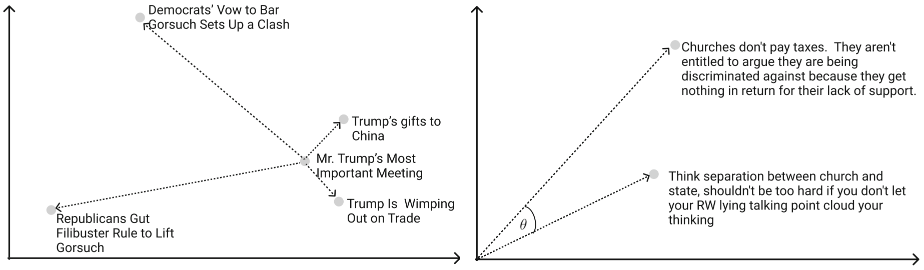

Democrats' Vow to Bar Gorsuch Sets Up a Clash
Trump's gifts to China
Mr. Trump's Most Important Meeting
Trump Is Wimping Out on Trade
Republicans Gut Filibuster Rule to Lift Gorsuch
Churches don't pay taxes. They aren't entitled to argue they are being discriminated against because they get nothing in return for their lack of support.
Think separation between church and state, shouldn't be too hard if you don't let your RW lying talking point cloud your thinking
θ

Fig. 1. Left figure: Find the most similar article with k-nearest neighbor. Right figure: Calculate the semantic similarity θ between the comment vectors. Comment and article source: footnote 4

In this last step, we then sort the comment candidate set based on the semantic similarity to our comment **C**. This is calculated using the cosine similarity of the vector space embeddings (Fig. 1 right). Here, we also calculate the average vector for the comments like we did for the keywords of the articles. Because the embedding process is very time consuming, we precompute the vector embeddings for all comments in our dataset. Finally, the n best comments with the highest semantic similarity are presented to the user in a ranked list.

4 Evaluation of the Model

In the following, we describe the evaluation of our approach for the comment-centric-comment recommendation. First, we explain the embedding models and the dataset we used for the experiment. Then, we describe the annotation tool we developed to assess the quality of the suggestions of the model. Finally, we present the results of our evaluation.

4.1 Embedding Model

During the development of our model, we noticed, like [11], that the choice of the embedding model used is important. Therefore, we have decided to test different pre-trained embedding models and compared the results to decide which embedding model is the best for our approach.

For this, we used the PyPi framework sentence-transformers[2][9] of the UKPLab. This package provides simple methods to generate vector representations of sentences and paragraphs, providing a large number of general purpose models trained on more than 1 billion training pairs.

From this set of models we selected five models and compared them with each other in the course of the evaluation:

- Paraphrase-MiniLM-L12-v2 (PML12)
- Bert-Large-Uncased (BLU)
- Paraphrase-Mpnet-Base-v2 (PMB)
- Stsb-Mpnet-Base-v2 (SMB)
- Stsb-Roberta-Base-v2 (SRB)

4.2 Dataset

To the best of our knowledge, there is no annotated dataset with comment suggestions for a give comment. For this reason, we searched for a promising dataset to extract comments from and asses these recommendation in an evaluation.

In this process, we came across a Kaggle dataset of comments from the New York Times[3]. The dataset consists of over 2 million comments from 9000 articles published in the New York Times from Jan-May 2017 and Jan-April 2018.

For the evaluation of the embedding models, fifteen comments from different topics like politics, trade etc. have been chosen. Six comments were extracted for each comment as proposals and were then evaluated by the annotators. We made sure to select both longer and shorter comments. This way, it was investigated whether the model could find good suggestions for both long and short comments. This was important because during the embedding process, as explained in Sect. 3, we generate an average vector representation for the comment that we want to find suggestions for. Therefore, the question is if short comments have enough information to find a suitable suggestions or if long comments have too much information so that the information overlap and the vector has no clear direction.

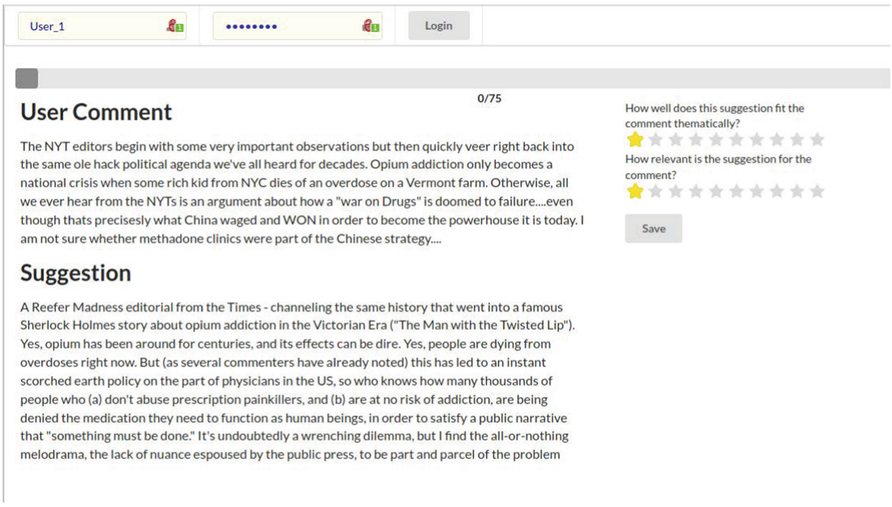
4.3 Annotation Tool

As mentioned in the previous section, there is no annotated dataset that allows us to evaluate our model and the different embedding models. Therefore, we

[2] https://pypi.org/project/sentence-transformers/.

[3] https://www.kaggle.com/aashita/nyt-comments accessed 09/21/2021.

evaluated the suggestions made by the respective embedding models with the help of four annotators from the field of computer science. In order to keep the annotation as simple as possible, we developed an annotation tool with an easy-to-use interface to help the annotators with assessing the quality of the suggestions (Fig. 2).



User Comment

The NYT editors begin with some very important observations but then quickly veer right back into the same ole hack political agenda we've all heard for decades. Opium addiction only becomes a national crisis when some rich kid from NYC dies of an overdose on a Vermont farm. Otherwise, all we ever hear from the NYTs is an argument about how a "war on Drugs" is doomed to failure....even though thats precisely what China waged and WON in order to become the powerhouse it is today. I am not sure whether methadone clinics were part of the Chinese strategy....

Suggestion

A Reefer Madness editorial from the Times - channeling the same history that went into a famous Sherlock Holmes story about opium addiction in the Victorian Era ("The Man with the Twisted Lip"). Yes, opium has been around for centuries, and its effects can be dire. Yes, people are dying from overdoses right now. But (as several commenters have already noted) this has led to an instant scorched earth policy on the part of physicians in the US, so who knows how many thousands of people who (a) don't abuse prescription painkillers, and (b) are at no risk of addiction, are being denied the medication they need to function as human beings, in order to satisfy a public narrative that "something must be done." It's undoubtedly a wrenching dilemma, but I find the all-or-nothing melodrama, the lack of nuance espoused by the public press, to be part and parcel of the problem

How well does this suggestion fit the comment thematically?
★☆☆☆☆☆☆☆☆☆

How relevant is the suggestion for the comment?
★☆☆☆☆☆☆☆☆☆

Fig. 2. Screenshot of the annotation tool. The annotators asses the thematic affiliation and the relevance of the suggestion for the given user comment. The annotators rates the suggestions on a 1 to 10 scale.

The annotators were presented with each of the six proposals per comment in a separate step, so that the proposals are first evaluated independently. For each proposal, the annotators are asked to rate on a scale of 1 to 10 whether the proposal fits thematically and whether it is a relevant proposal. “Relevant” means that the suggestion offers new insights, perspectives, or information for the topic. Following, the annotators are presented with the six suggestions as a ranked list, with the most relevant comment at the top. They are then asked to rate the order of comments. Does the ranking fit the relevance of the comments? This results in a total of 525 pairs of annotations for all five models, which every annotator has to evaluate.

To prevent a bias in the evaluation, the comments and models are presented to the annotator in a random order as packages. This means that the six suggestions followed by the ranked list of suggestions for a comment were presented to the annotator as a contiguous package and then a new comment with six suggestions and a ranked list is selected randomly.

4.4 Evaluation

Model Performance. One of the most important qualities of a model for real-time comment suggestion, besides the quality of the suggestion, is of course the performance of the model. A model that cannot generate the suggestions in feasible time would defeat its purpose to help the user in an online discussion because online discussion move at a fast pace.

For this reason, in addition to annotating the proposals, we also have measured the average time our model needs to find a suggestion for a given comment per model. The experiment was conducted on a Lenovo ThinkPad-T15p-Gen-1 with an Intel Core i5 with 8x2.60GHz and 64 GB RAM:

Table 1. Time measurement for extracting one proposal per model

Modelname	Time (s)
stsb-mpnet-base-v2	1.2
bert-large-uncased	1.9
paraphrase-mpnet-base-v2	1.3
stsb-roberta-base-v2	1.3
paraphrase-MiniLM-L12-v2	1.0

Suggestion Quality. In the following paragraph, we evaluate the quality of the suggestions for the different embedding models. As explained in Sect. 4.3, we have three different parameter to assess our model. First, we evaluate the six suggestions $\mathbf{S_1}, ..., \mathbf{S_6}$ for every comment $\mathbf{C}$ the user is interested in seperatly. Here, we rate two quality dimensions. On the one hand, we rate how good the suggestion fits thematically to the comment $\mathbf{C}$ and on the other hand, we evaluate how relevant the proposal $\mathbf{S_i}$ is. We rate these two dimensions because a suggestion, which is not at all or hardly connected to the comment the user is interested in, cannot be relevant for the user. However, only because a suggestion fits thematically to $\mathbf{C}$, it does not mean it is also relevant for the user. For example, if the comment the user is interested in reads like this:

> *I relate to this completely. I remedied the situation by getting off Facebook completely and I do not miss it one bit. I used to log on to my computer and without thinking type Facebook.com but no more. I am so relieved it's incredible how much people are sucked into that. No more distant people I wish were gone than lingering. I have whatsapp notifications muted it's on my time I check my msgs. I highly recommend you utilize the tech in your phone and get yourself some much needed privacy. Cheers*[4]

A suggestion like the following would fit thematically, but is hardly relevant for the user because the suggestion does not contribute any relevant information to the discussion or shows new insights:

[4] https://www.kaggle.com/aashita/nyt-comments accessed 09/21/2021

Wow! Instagram has really boosted my self-esteem and provided long-term happiness! Said no kid, ever.[5]

As a final parameter, we looked at the ranking of the suggestions by the semantic similarity. With this parameter, we investigate if the semantic similarity is sufficient to be the foundation for the extraction of relevant suggestions or whether further parameters need to be used in future experiments.

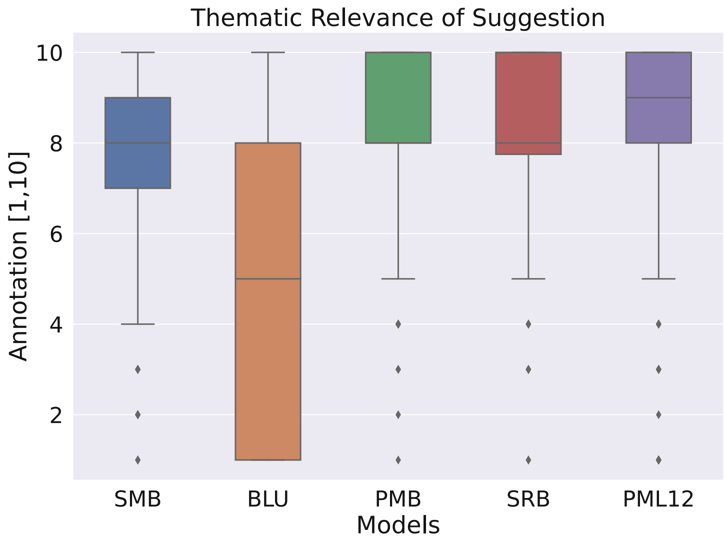


Fig. 3. Boxplot with the scores to asses how good the different embedding models extract suggestions that fit thematically to the comment the user is interested in; higher scores are better

In Fig. 3, we see the results of the evaluation for the quality dimension of how good the respective suggestions $\mathbf{S_1}, ..., \mathbf{S_6}$ fit thematically to the comment $\mathbf{C}$ the user is interested in as boxplots. We notice that all models except for Bert-Large-Uncased (BLU) produce very good results. The median for stsb-mpnet-base-v2 (SMB), paraphrase-mpnet-base-v2 (PMB), and stsb-roberta-base-v2 (SRB) lies at 8 and 75% of the scores are above 7 respectively 8. Only Paraphrase-MiniLM-L12-v2 (PML12) produces higher results with a median of 9 and 75% of the scores above 8. Therefore, we can state that these four embedding modes produce good and reliable results. In contrast to this, the median of Bert-Large-Uncased (BLU) lies at 5 and the scores scatter very much from 1 to 8. Therefore, the suggestions extracted with Bert have a high variance and do not produce a reliable quality. We observe that with the other four embedding models, we can extract suggestions from the dataset that fit thematically much better and are more consistent in the quality than these extracted with Bert.

[5] See footnote 4

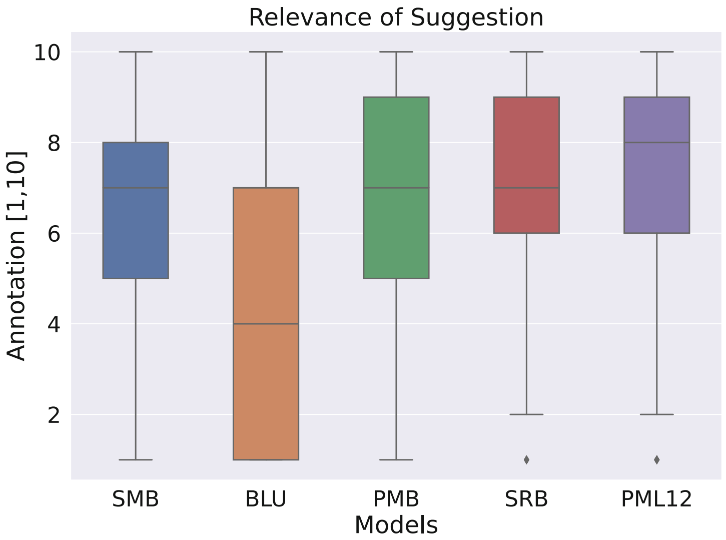


Fig. 4. Boxplot with the scores to asses how good the different embedding models extract relevant suggestions for the comment the user is interested in

Figure 4 shows a similar picture as Fig. 3. Again Stsb-Mpnet-Base-v2 (SMB), Paraphrase-MiniLM-L12-v2 (PML12), and Stsb-Roberta-Base-v2 (SRB) have the same median. Here with a value 7 with 75% of the scores above 5 respectively 6. Paraphrase-MiniLM-L12-v2 (PML12) has the highest median of all models with a value of 8 with first quartile at the score of 6. The worst results are, again, these of Bert-Large-Uncased (BLU). The median for it is 4 and the first quartile lies at 1 and the third quartile at 7. Therefore, the results scatter again very strongly and are much worse and much less reliable as the results of the other models. However, we must note that for the relevance the other models scatter more as for the thematic affiliation. This is not surprising since it is much easier to determine the thematic affiliation by the semantic similarity as the relevance of a suggestion. However, the relevance represents the more important quality dimension for our model, since the thematic affiliation is only a prerequisite for the relevance, but does not add any value on its own as we can see in the quote at the beginning of the Sect. 4.4.

Figure 5 shows the last parameter we have examined in our evaluation, the ranking of the suggestions. We notice the same trend as in the figures above that Bert-Large-Uncased (BLU) produces the weakest results. However, the results of the different embedding models are much closer to each other than above. Here, Paraphrase-MiniLM-L12-v2 (PML12), Stsb-Roberta-Base-v2 (SRB), and Paraphrase-Mpnet-Base-v2 (PMB) have a median of 7 with the first quartile at 5 respectively 6. Stsb-Mpnet-Base-v2 (SMB) has a median of 6 with 75% of the scores above a value of 5. The worst results are, again, produced by Bert. Here, we have a median of 5 with the first quartile at 3. Again, we can state that the

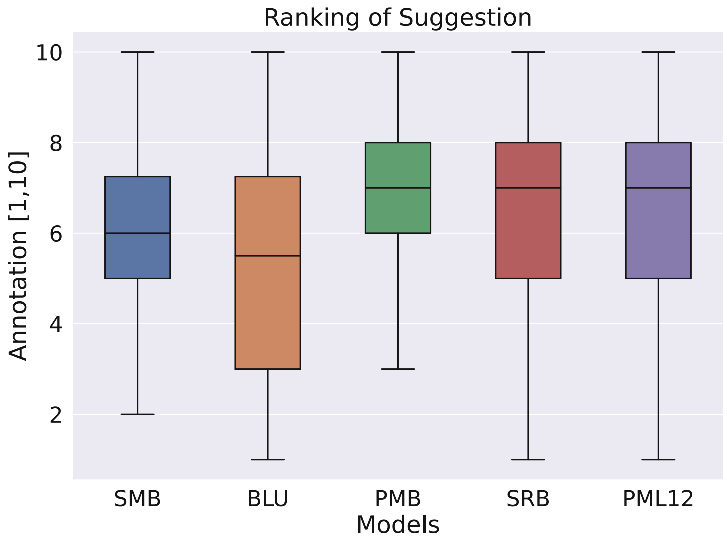


Fig. 5. Boxplot with the scores to asses how good the different embedding models rank the suggestions for the comment the user is interested in

other embedding models produce better results with more consistency. However, the distance to Bert here is much closer than before. Overall, we can see here that the ranking results are good but improvable. This indicates that the semantic similarity is a good starting point for our approach and this model provides a solid baseline for future experiments. Nevertheless, we need to consider other parameters for future models to improve these results.

Model Performance for Long and Short Comments. As explained in Sect. 4.2, we selected both short and long comments to investigate how the model performs for comments of various length because short comments might not contain enough information to find suitable suggestions for and long comments could contain to much information so they might overlap in the process of calculating the average vector for the comment.

Table 2 shows the average scores for the three quality dimension where we considered a comment as short if it contains less that 50 words:

Table 2. Average scores for the different quality dimensions for short and long comments

Model name	Thematic (short/long)	Relevance (short/long)	Ranking (short/long)
stsb-mpnet-base-v2	(6/6)	(8/7)	(6/6)
Bert-large-uncased	(6/7)	(8/8)	(5/6)
paraphrase-mpnet-base-v2	(3/4)	(4/5)	(4/5)
stsb-roberta-base-v2	(6/7)	(8/8)	(6/6)
paraphrase-MiniLM-L12-v2	(7/6)	(8/8)	(6/6)

Scores for short and long comments only differ by a score of 1 point on a scale from 1 to 10, if they differ at all. Therefor, we can affirm that the model produces consistent results for all embedding model despite the length of the comment.

Inter-Annotator-Agreement. The evaluation of the suggestions for a comment is a very subjective annotation task. Some users might consider a suggestion very relevant for the comment the user is interested in, while others do not. In many cases this depends on the background knowledge and experiences of the user. Therefore, it is very important to have a quantitative value to determine the agreement among the annotators. Otherwise it is hard to interpret the results from Sect. 4.4 because it would be possible that the scattering of the boxplots results come from the fact that the annotators are divided in their opinion about the model and some asses the quality of the model very poorly while others very good. However, if the annotators are very consistent in their judgment of the model then we can explain the scattering as an inconsistency in the quality of the suggestions the model produces.

For this reason, we have calculated the Inter-Annotator-Agreement (IAA) of the annotators. Due to the high subjectivity of the annotation task, we have decided to calculate the IAA for the quality dimension if the annotators considered the proposal for a comment as good and calculated the percentage of annotations with a score ≥6:

Table 3. Proportion of annotation values greater than or equal to 6

Modelname	Thematic	Relevance	Ranking
stsb-mpnet-base-v2	0.91	0.71	0.72
Bert-large-uncased	0.49	0.41	0.50
paraphrase-mpnet-base-v2	0.95	0.75	**0.77**
stsb-roberta-base-v2	**0.96**	**0.78**	0.68
paraphrase-MiniLM-L12-v2	0.93	0.77	0.70

Besides allowing us to calculate an Inter-Annotator-Agreement, it also gives us an quantifiable value for the quality of the suggestions produced by the embedding models. Looking at Table 3, we see that the model Stsb-Roberta-Base-v2 extracts the suggestions that fit thematically the best and are most relevant, while Paraphrase-Mpnet-Base-v2 provides the best ranking. However, one also finds that the difference with Stsb-Mpnet-Base-v2, Paraphrase-Mpnet-Base-v2, and Paraphrase-MiniLM-L12-v2 is insignificant. Not so with Bert-Large-Uncased. Here, the results are much worse than for the other embedding models. This reflects more or the less the results from Sect. 4.4 where Bert provided the worst results and the other embedding models are very close to each other.

To then calculate the Inter-Annotator-Agreement, we have used Krippendorf Alpha (α)[6][5]. We achieved an α of **0.46** for the relevance, **0.70** for the thematical relevance, and **0.69** for the ranking. Thus, there is a high agreement between the annotators despite the high subjectivity of the annotation task.

Self-Agreement of Annotators. It happens from time to time that some models propose the same suggestion for a comment. Furthermore, this is a very laborious annotation task that the annotators have performed over a few weeks. To ensure the consistency of the annotation, we also examined how much the annotators agreed with themselves over the course of the annotation. For this purpose, the mean of the deviation of all annotators with themself was calculated. It only makes sense to look at the thematic agreement and the relevance, since no two models produced exactly the same sequence for the ranked list of suggestions. For the thematic agreement, we have an average deviation of **0.53** and for the relevance of **0.7**[7]. This indicates a high self-agreement among the annotators because on average the difference in the scores for the same proposal made by a different embedding model is below **1**.

4.5 Discussion

After reviewing the results of our evaluation, we can state that our model produces good and consistent results for proposing suggestions for the comment the user is interested in. Furthermore, we have a high agreement among our annotators with and among themselves which means that despite having a very subjective annotation task the annotators agree in their assessment of the model.

However, we can not definitively determine why Bert produces the worst results in all quality dimension we have considered, we assume that in the process of generating the average vector representation of the comment we loose information of the content of the comment because Bert produces very context-sensitive vector representation.

The results of the model depend on which embedding model one chooses. Many publications in the field of argument search and the like use Bert as the embedding model of choice [3,6,10], since it can produce excellent embeddings through the use of context. However, for our problem BERT was not the model of choice.

5 Conclusion and Future Work

In our work, we introduced a new approach to the field of comment recommendation and presented our model that will serve as a baseline for future experiments. In contrast to previous work, we shift the focus away from the user to the comment the user is interested in. By this, we try to offer new perspectives and

[6] Range: -1 (perfect disagreement) to 1 (perfect agreement).

[7] Range: 0 (perfect agreement) to 10 (perfect disagreement).

insights to the user. To do this, we made use of a similar approach like [3] to create a candidate set of comments and select the six most interesting suggestions from it by their semantic similarity to the comment. For this, we use very basic machine learning methods like k-nearest-neighbor and vector embeddings. Due to the lack of an annotated dataset to evaluate our results, we asked four annotators to assess the quality of the suggestions for the three parameter. How does the suggestion fits the comment thematically? How relevant is the suggestion to the comment? How good is the ranking of the six suggestions for the comment? Taking into account the simplicity of the model and that we have a very subjective annotation task, we can conclude that our model serves as a very good baseline for future experiments and is a good approach for comment-centric-comment recommendation. We also note that semantic similarity is a good start for our model, but is not sufficient as the only parameter for the model and we need to consider other in future experiments.

In future work we will continue our work in two directions. First, we want to transition our model into an ecosystem of tools that will allow us to test our approach and model in real world discussions under news articles. For this, we develop a browser plugin that allows the user to get suggestions for every possible comment under any news article and a web scrapper to keep our knowledge base up-to-date by retrieving comments from discussion of different news agencies. Second, we want to try different approaches to improve upon our baseline model and to ensure a well balanced selection of comments. If we want to make a significant contribution to online debates and to combat the echo-chamber, we have to put emphasize on a selection of suggestions that represents different views and political orientations on a topic.

Acknowledgment. We would like to thank Markus Brenneis and Björn Ebbinghaus for their help with the annotations.

References

1. An, J., Kwak, H., Posegga, O., Jungherr, A.: Political discussions in homogeneous and cross-cutting communication spaces. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, pp. 68–79 (2019)
2. Del Vicario, M., Zollo, F., Caldarelli, G., Scala, A., Quattrociocchi, W.: Mapping social dynamics on Facebook: The Brexit debate. Social Netw. **50**, 6–16 (2017)
3. Dumani, L., Neumann, P.J., Schenkel, R.: A framework for argument retrieval: Ranking argument clusters by frequency and specificity. Adv. Inf. Retrieval **12035**, 431 (2020)
4. Flaxman, S., Goel, S., Rao, J.M.: Filter bubbles, echo chambers, and online news consumption. Public Opinion Q. **80**(S1), 298–320 (2016)
5. Krippendorff, K.: Computing krippendorff's alpha-reliability (2011)
6. Ollinger, S., Dumani, L., Sahitaj, P., Bergmann, R., Schenkel, R.: Same side stance classification task: Facilitating argument stance classification by fine-tuning a Bert model. arXiv preprint arXiv:2004.11163 (2020)
7. Park, D., Sachar, S., Diakopoulos, N., Elmqvist, N.: Supporting comment moderators in identifying high quality online news comments. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 1114–1125 (2016)

8. Quattrociocchi, W., Scala, A., Sunstein, C.R.: Echo chambers on Facebook. Available at SSRN 2795110 (2016)
9. Reimers, N., Gurevych, I.: Sentence-Bert: Sentence embeddings using Siamese Bert-networks. arXiv preprint arXiv:1908.10084 (2019)
10. Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., Gurevych, I.: Classification and clustering of arguments with contextualized word embeddings. arXiv preprint arXiv:1906.09821 (2019)
11. Schaefer, R., Stede, M.: Improving implicit stance classification in tweets using word and sentence embeddings. In: Benzmüller, C., Stuckenschmidt, H. (eds.) KI 2019. LNCS (LNAI), vol. 11793, pp. 299–307. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30179-8_26
12. Wang, Y.: Comment section personalization: Algorithmic, interface, and interaction design. In: Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pp. 84–88 (2021)
13. Zhou, M., Shi, R., Xu, Z., He, Y., Zhou, Y., Lan, L.: Design of personalized news comments recommendation system. In: Zhang, C., et al. (eds.) ICDS 2015. LNCS, vol. 9208, pp. 1–5. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24474-7_1

Chapter 4

Building Tailored Comment-Recommendation Prototypes Using a Modular Design Framework

In this chapter, we give an overview about the contributions and impact of our paper Steimann and Mauve (2024):

Jan Steimann and Martin Mauve.

"Developing Custom-Made Comment-Recommendation Prototypes with a Modular Design Framework"

In: *Proceedings of the 16th International Conference on Social Computing and Social Media*, held as part of the 26th HCI International Conference, pages 97–112, Springer International Publishing
Acceptance Rate: ∼55.2%

4.1 Summary

This paper presents our open-source software framework designed to improve the development of comment-recommendation prototypes. The Introduction outlines the importance and potential applications of such a system. With the vast influx of comments in online sections, users can quickly become overwhelmed and struggle to maintain an overview. Researchers aim to assist users through comment recommendation systems that pinpoint, e.g., insightful views for the user like Zhou et al. (2015b), as seen in the related work. However, these publications often emphasize model development and provide evaluations primarily in artificial settings. Rarely are these models tested in real-world scenarios with actual users, due to the demanding nature of such experiments in terms of time and resources. The software required to perform real-world evaluations is particularly expensive to develop. This publication seeks to ease this gap by offering an accessible framework that provides the technical infrastructure needed for real-world prototyping, allowing researchers to focus solely on implementing the recommendation models they wish to test.

Section 3 of the paper outlines the core concept of the framework, highlighting features that streamline the creation of prototypes catered to researchers' unique requirements. The framework provides an initial prototype where researchers only need to incorporate the recommendation model they wish to assess. Its modular architecture stands out, allowing effortless updates or replacements of components to align with the experiment's demands. Additionally, a built-in web scraper facilitates the collection of essential data for experiments. These features are thoroughly examined in the paper. Section 4 of the paper delves into the framework's implementation, followed by two example applications that illustrate its practical use.

4.2 Personal Contribution

Jan Steimann was the principal contributor to this paper. He devised the concept and carried out the implementation of the open-source software framework. In addition, he performed the two example implementations and authored the paper. Martin Mauve initially proposed the idea of expanding the original prototype of a comment recommendation system into a comprehensive framework. He also offered feedback and engaged in discussions regarding the concepts.

4.3 Importance and Impact on this Thesis

This paper presents an intuitive open-source framework designed for the rapid development of comment recommendation systems. It provides the essential infrastructure for building and extending our model described in Chapter 6 and for future endeavors. This framework enables the creation of recommendation systems that are straightforward to test in real-world scenarios and can be tailored for specific purposes. The model detailed in Chapter 6 was evaluated using an evaluation app that accesses recommendations directly via the model's REST API, demonstrating the framework's versatility in custom applications—a significant advantage. This flexibility is one of our key contributions, alongside the ready-to-use prototype. The framework offers researchers a comprehensive technical base that can both function independently with the given user interface and be incorporated into other systems.

Developing Custom-Made Comment-Recommendation Prototypes with a Modular Design Framework

Jan Steimann( ) and Martin Mauve

Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany
jan.steimann@hhu.de

Abstract. Comment sections of news articles are a popular way to discuss the contents of these articles. But the number of comments posted every day has become so large that almost no one can get a solid overview about the discussion. To address this problem, there are many approaches for comment recommendation systems. However, they tend to focus mostly on the development of sophisticated models to combat this problem while evaluating their systems in limited and mostly artificial settings. In our paper, we introduce a modular open-source software framework for the development of comment recommendation prototypes that can be used to evaluate models in real-world environments. The modularity allows developing systems that are adapted exactly to the use-case or model one needs. This concept allows exchanging and adapting the different components of the concept to test e.g. different user-interfaces or recommendation models. To show the usability of our framework we present the implementations of two comment-recommendation applications.

Keywords: Information Retrieval · User-Interface · Software Development Framework · Recommendation System

1 Introduction

Comments are a popular tool for the exchange of views in online discussions of news media articles. They offer readers the opportunity to exchange opinions about the content of articles and to discuss different points of view. However, the number of comments posted every day has become so large that it is hard – if not impossible – to get a good overview of the discussion. Articles about controversial topics like booster shots for COVID-19 can have a few hundred comments like [7] or even a few thousand like articles about Russia's president Putin [13]. To aid users in navigating the flood of comments posted every day, there exist various approaches, which we will take a look at in Sect. 2.

Many of these approaches so far have focused primarily on developing sophisticated models and only present prototypes in mostly artificial settings. However,

A. Coman and S. Vasilache (Eds.): HCII 2024, LNCS 14703, pp. 97–112, 2024.
https://doi.org/10.1007/978-3-031-61281-7_7

to realistically assess a model's effectiveness, it is equally important to test it in authentic scenarios. Therefore, it is crucial to evaluate these models under real-world conditions to obtain a genuine impression of their capabilities.

In this paper, we present a solution to this problem by introducing our open-source software framework for developing practical comment recommendation prototypes. These prototypes can be customized for various real-world applications while at the same time reduce the development effort for the evaluation of recommendation models. For this, we provide a complete infrastructure with several exchangeable components. This infrastructure consists among other things of a front-end user interface and backend server. Adjustments to the different components can be easily done without making major changes to the rest of the system. Our new framework reduces the development effort for testing e.g. a new user-interface or recommendation model.

In the following section, we take a look at previous work in the field of comment recommendation. In the third section, we elaborate on the concept for our comment recommendation system and in the fourth section, we explain the implementation of the framework based on the concept of section three. In the last section, we present two example implementations to demonstrate the usefulness of the framework to speed up the development of comment recommendation prototypes that can be used in real-world scenarios.

2 Related Work

Recommendation systems for comment sections of news article have been a popular research field for quite some time since this is an important field of the public debate. These systems can help the reader to get a well-informed opinion by suggesting articles and comments that cover topics which are of great interest for the user. For example, [14] uses the comments of the discussion in which the user currently participates to suggest articles that may interest them. The authors of [11] try a similar strategy by recommending news stories that the user probably will comment.

The recommendation of articles and news stories that the user finds especially interesting is only the first step in supporting the public debate. The actual participation in the public debate begins when the user takes an active part in the discussion and reads comments of other users and formulates their own contribution. Since the beginning of online discussions, various problems have emerged, which are partly induced by the medium itself.

One problem is the vast amount of comments which are posted under the articles. Some article may have few hundred or even a few thousand comments if they cover a current and very controversial topic. For this reason, there are different approaches to help the user navigate the flood of comments. For example, [15] uses a personalized comment recommendation system which sorts out low quality comments and classifies the rest in *insightful view* or *informal comment*. At the end, the system recommends comments to the user based on previous posts from the user. Another approach [5] tries to help the user navigate large discussions by using topic modeling. The system extracts an initial set of topics for the discussion and improves them with on-the-fly user feedback. The topics are then visualized in a user-interface that should help the user to get a better overview of the discussion and quickly identify insightful views in the debate.

Another major problem of online discussion in general is the creation of echo-chambers and filter bubbles. Here, users often find a very homogeneous set of opinions and thus are rarely confronted with ideas that are contrary to their worldview. As a result, these opinions become entrenched, and when users with different opinions clash on a controversial issue such as climate change or migration, a constructive debate is hardly possible.

To address this issue, there are various research approaches, which all have in common that they try to present the user a broader spectrum of opinions during the opinion forming process. In [2], the authors use various techniques like dimensionality reduction with PCA, collaborative filtering, etc. and then present the user an interface that automatically highlights comments that many other users found particularly insightful. The authors of [4] try to help the user to reflect on their own position on the topic by presenting comments which oppose their own stance for the topic and mark them as *recommended*. Comments that align with the view of the user are marked as *not-recommended*.

2.1 Ethical Considerations, Impact, and Intended Application

All these approaches have in common that they mainly focus on the development of algorithms for recommending comments and articles to the user. However, often the evaluation under real-world conditions is not emphasized as much and only demonstrations in a mostly artificial settings are presented. Some work even presents only a design idea how an application for the presented recommendation algorithm might look like. Nevertheless, we believe that the development of prototypes to test the algorithm in a real-world scenario is as important as the development of even more sophisticated algorithms. Therefore, we want to bridge the gap between theoretical research and real-world applications with this paper by presenting an open-source framework for the development of such prototypes.

This framework will help researchers to speed up the evaluation of recommendation models by offering a customizable ready-to-use prototype where only the algorithm that should be tested has to be injected. The researcher can use the prototype for the early development and first internal tests and afterwards customize the prototype, e.g. the user-interface, for further evaluation in larger user studies. Additionally, the framework offers a complete web scraper that only has to be configured for the website that should be included as a data source. Alternatively, the researcher can import any CSV dataset into the database with the provided script.

Our work shares the same ethical implications as all work on recommendation systems in general. It can be used, intentionally or unintentionally, to develop systems that shape certain user behavior in a mischievous way. For example to reinforce opinions about a topic by presenting only article and comments to the user that support a certain opinion. Preventing intentional mischievous behavior is out of scope for our work. However, we believe that our system can help researchers to rigorously test their recommendation algorithm as realistic as possible and by this identify potential negative implications.

3 Comment Recommendation Framework Concept

In this chapter we will explain the concept and structure of our comment recommendation framework and the individual components. We expound on the details of the components and how they interact with each other. In the following, we start from a high level design point of view and then outline the details.

3.1 High Level Design

In the following subsection, we take a look at the contribution of our concept from a high level.

Framework. The main contribution of our framework is the rapid prototyping of comment recommendation systems through our ready-to-use system. The researchers only have to implement the model they would like to test and populate the database either with the news agency scraper or importing a dataset via the CSV reader component which we will both explain in Sect. 3.3. Another advantage of our system is that it can be used in both, a controlled lab scenario as well a real-world scenario, thus providing the researcher with a vast variety of evaluation possibilities. This provides a huge advantage compared to the previous approaches, we have examined in Sect. 2, which often present demonstrations in mostly artificial settings. Furthermore some approaches only present ideas or concepts of how their system could be used in a real-world scenario because the development of a prototype is very time-consuming. Especially, these works would benefit of our framework.

Interchangeable Components. We developed the recommendation framework with a modular design concept in mind. Therefore, the framework consists of several components that can replaced or updated to a new use-case if needed. This allows researches to change or replace every component of the prototype to adapt it to their use-case or to test the effect of changes e.g. in the user-interface or model. This way, they can quickly test new ideas or models without developing a new prototype. Instead, they only replace or change a specific component.

Data Acquisition. A key point of a comment recommendation system is the data and how they are acquired. We provide two solutions for this problem with our framework. If the researches already have a dataset and would like to use it for the prototype, then we provide the CSV reader component to import the dataset into the database of the recommendation prototype. However, if the researchers would like to create a new data basis for their prototype, we offer a news agency scraper system that the researchers can use with little configuration to scrape different news agency sites.

3.2 System Overview

Next, we examine the system in its entirety. As explained in Sect. 3.1, the system consists of several components that were developed independently of each other and are intended to fulfill different roles in the framework. We developed the concept in a modular design web application approach so that every component can easily be replaced or extended. This way, the comment recommendation system should be adaptable for various purposes and we can use the prototype in a laboratory and also in a real-world setting.

If we take a look at the system in its overall form in Fig. 1 we see that the system can be divided into two main parts, front-end and backend.

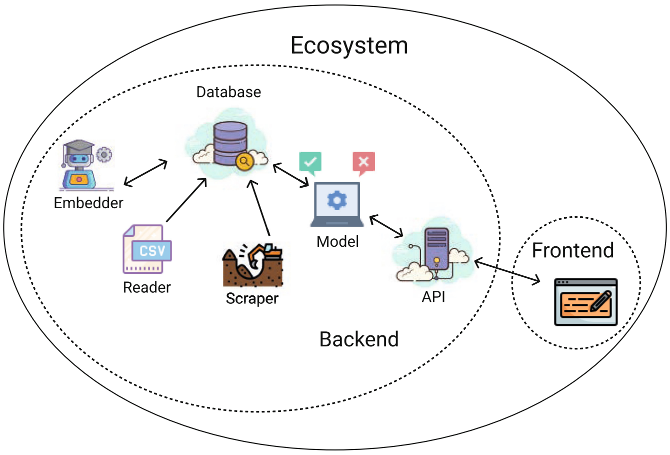


Fig. 1. Overview of the ecosystem concept and its different components. Icons by [6]

On the one hand, we have the front-end, which handles the direct interaction with the user and presents them with the results of the recommendation process.

When the user triggers the recommendation process, a request is sent to the backend server. The backend server receives and extracts all necessary information for the recommendation model from the request. Subsequently, the model is triggered with this information to start the retrieval process. For this purpose, the model makes several queries to the database to gradually extract the final result set, which is then transferred back to the front-end component to render the recommendations.

The database, from which the suggestions have been extracted, builds the foundation for the entire recommendation process. All suggestions that can be presented to the user are extracted here. Having a sufficient data basis is therefore of decisive importance for presenting suitable suggestions. For this purpose, the database is regularly populated with the help of a comment scrapper. The scraper can be configured to query different news agency websites and collect the comments of the current articles and save them in the database. Additionally, if the user possess a dataset they would like to use, our framework provides a CSV reader component to import the dataset into the database. Afterwards, we have to trigger the embedder which calculates the embedding vectors of the new comments and articles with a machine learning model and stores them in the database.

3.3 Components in Detail

In the following, we take a look at the individual components in detail and examine what function they fulfill within the framework.

Frontend. First, we take a look at the front-end component. Depending on the purpose of the system, different user-interfaces can be implemented here to help the reader to interact with comment sections on various websites.

When the user triggers the recommendation process, the UI sends a request with the informations that are needed by the recommendation model to extract suitable suggestions from the database. Afterwards the user-interface waits for the corresponding response from the server to render the list of suggested comments for the user.

Due to the separation of UI and backend communication, it is also easy to replace the complete user-interface with a different approach and to integrate the comment recommendation system in a larger web-system.

API. After the request has been sent by the front-end, the backend server takes over by processing and extracting all necessary information for the recommendation model from the request. The server then hands over the information to the model and afterwards receives the recommendations. Finally, the backend transmits the generated suggestions as a response back to the front-end. Therefore, the backend serves as the interface between the front-end and the recommendation engine.

We choose to use an interface structure so that the recommendation model can be replaced very easily by another model, depending on the purpose of the application prototype. To use another model, we only have to adapt the information extraction from the request.

Model. The model acts as the core of the system. It uses the informations the reader transmits to retrieve a selection of comments from the database.

For the framework, it is irrelevant which model we use. Depending on the use case of the system a different model can be injected into the system at any time if the model uses the same information as before. If the model needs a different set of information for the retrieval process, then these have to be part of the request from the front-end and the API needs to updated accordingly.

When the model is integrated in the code, it is only important that the model provides the correct interface method. How the model extracts the suggestions from the database is of no concern for the rest of the system. This way, it is easy to replace the model.

Database. For the database, we use a graph database which has the great advantage that the connections of the data are first level citizens. This means that we do not have to create the connections between our data by using joins, like relational databases, but the connections are already stored in the database like the nodes. Therefore, the retrieval process can run faster. This is essential for a system that is used in a real-time scenario like an online discussion. If the user has to wait several minutes for the recommendations they will most likely do not use the system.

News Agency Scraper. To present a suitable selection of comments to the user, the recommendation model needs a solid data basis and therefore, we need a system to populate the database. For this, we offer a comment scraper that allows the researcher to scrape arbitrary news agencies that use comment sections. It retrieves the comments from the articles and stores them in the database according to the database schema the recommendation model needs.

CSV Reader. If the researchers already have a dataset they would like to use instead or as an addition to the news agency scraper, we provide the CSV reader component that imports the given dataset into the database.

Embedder. In order for the recommendation model to work with the comments and articles in the database, the text must be in a format that can be processed by the model. For this reason, the last component needed for the framework is an embedder. Here, the researchers implement the embedding method that provides the embeddings that are needed for their recommendation model. We provide a script that fetches the new comments and articles that were just stored in the database by the scraper and calculates the embeddings for the texts. Afterwards, the nodes in the database are updated with the new embeddings.

4 Implementation of the Framework

In the following, we will examine the implementation[1] of the concept presented in Sect. 3. The system was developed in such a way that every component can

[1] https://github.com/hhucn/Comment-Recommendation-Framework.

be easily extended or replaced. By this, the implementation can be adapted to various use-cases. Here we take a look at how the concept has been implemented and how to utilize the framework to help researches with rapid prototype development. In Fig. 2, we see an overview of the framework with the concrete tools and libraries we use.

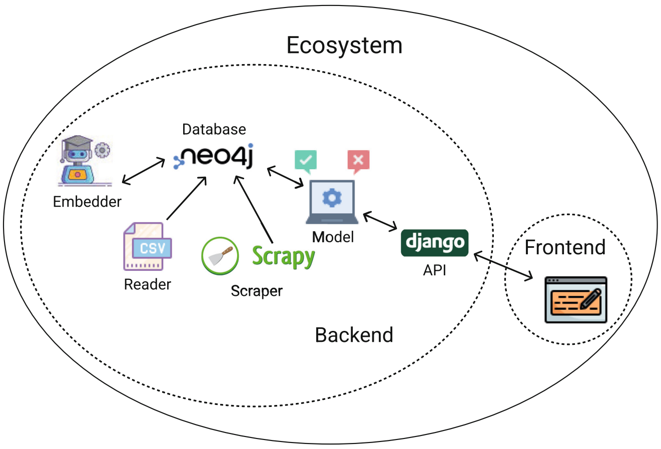


Fig. 2. Overview of the implementation of our framework. Icons by [6]

4.1 Components

In the following, we take a look at how the individual components have been implemented.

Front-End Component. For the user interface we decided to use a Chrome Extension with React [8]. A Chrome Extension is a very small application that can be installed in any Chrome Browser. This gives the researchers various possibilities for their experiment. On the one hand, they can provide a controlled laboratory scenario with pre-configured computers with the Chrome Extension already installed. Or on the other hand, the participants of their study can install the Chrome Extension on their personal computer at home. This opens more opportunities for the researcher to conduct their experiments.

The extension interacts with the website of the news agency and extracts the necessary information for the recommendation model. It then sends these information to the backend server and renders the response with the recommended comments to the user.

The advantage of a Chrome Extension is that it is build modular and we can replace the individual components of the extension at any time. Due to the modularity, we can update or replace different aspects of our user-interface with minimal changes on the rest of the components.

API Component. For the API, we use the Django framework [3] to process the requests from the front-end component. One advantage of Django is that we need very little configuration to process the requests. We only need one view where we instantiate our recommendation model and extract the information for the model from the request. Afterwards the model is called to retrieve the selection of comments from the database and finally the selection is packed in a JSON-response and send to the front-end component.

Database Component. For the implementation of the database explained in Sect. 3.3, we use a Neo4j graph database [9]. We choose a Neo4J graph database because it offers various advantages for our recommendation system prototype. As explained in Sect. 3.3, one advantage is that the connection between the nodes are already stored in the database and do not have to be queried. This speeds up the retrieval process if we need to connect a lot of nodes in the database.

The database schema for our implementation consists of two kinds of nodes. On the one hand, we have the comment nodes where we store the text of the comment and later the embedding vector for the comment text as fields in the node. On the other hand, we have the article nodes where we store the title, keywords, publication date, URL of the article, and the embedding for the article. Every comment that appeared in the comment section of an article is stored as a comment node and is connected to the corresponding article node as we can see in Fig. 3.

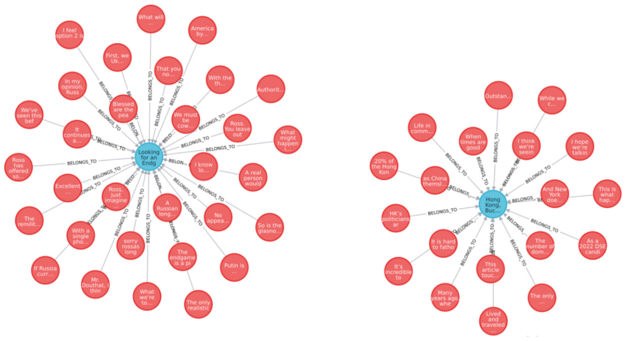


Fig. 3. Extract from the graph database. The blue nodes are the articles and the red ones are the corresponding comment nodes. (Color figure online)

This is the default configuration for the database nodes that is most likely sufficient for the majority of the prototypes. However, the system can be easily adapted with different attributes or even new node types as we will see in Sect. 5.2.

Recommendation Model Component. For the model component, we provide an abstract super class which specifies the interface method that all models have to implement. This method defines that it receives a dictionary with the data extracted from the request and returns a list of comments to be rendered by the user interface. Due to this, we can very easily replace the model with arbitrarily more complex models with only minimal changes to the system. Our new model just needs to extract the information needed from the dictionary and return a list of comments. It is completely irrelevant for the framework and the prototype how much complexity is behind the new model because of the interface method from the abstract super class.

News Agency Scraper Component. As explained in Sect. 3.3, our framework uses a comment scraper that can be configured to scrape arbitrary news sites to populate the database of Sect. 4.1.

For every news site that we want to scrape, we just need to create a new scraper component with functions specific to the HTML structure of the website.

For our implementation, we chose to use the Scrapy Framework [1] to scrape the websites. Our scraper component always works on the same principle. It uses the super class with the code for all news agency specific classes and makes a request to the website of the corresponding news agency for a list of articles. This list is then processed to extract all necessary information such as title, URL, keywords, and publication date. Normally, the comments are loaded in a second request for every article. This request is then also processed to get all comments belonging to the respective article. Finally, all these informations are stored accordingly to the database schema explained in Sect. 4.1.

CSV Reader Component. If the researchers have a dataset they want to use for their prototype and experiment, we also provide a CSV Reader component to import the data into the database. This can also be used as a supplement to the news agency scraper component to provide a sufficient data basis.

Embedder Component. After the News Agency Scraper or CSV reader is done with retrieving or importing articles and comments, we use the Embedder Component written as a Python script to compute the embeddings. The embedder calculates the representations the recommendation model from Sect. 4.1 needs to find the recommendations. For this, it queries all comments and articles from the database that do not have an embedding and computes a vector representation of the information the models needs with the embedding model the user implements.

4.2 Utilization of the Framework

In the following section, we examine how the framework can be utilized to develop custom-made comment recommendation systems.

Package. We implement the framework as a Python package[2] which allows the easy development of recommendation system prototypes. The user just needs to install the package and execute it. Afterwards, the package asks different questions with a dialog to determine which components of the package are needed for the prototype the user would like to develop and then creates a folder only containing the components needed. For example, if the user already has a dataset with articles and comments they would like to use for their experiment, the news agency scraper is obsolete and does not need to be part of the system. Instead, the CSV reader component is used.

Docker and Docker-Compose. To simplify the usage of the prototype in the development and experiments, we dockerized[3] the system and provide different docker-compose files[4] for the components of the system. This allows to run the recommendation system in an isolated environment and by this an easy portability and installation on different systems. Only Docker and Docker-Compose must be installed there. The necessary libraries and images will be downloaded and installed by Docker.

5 Example Implementations

In the following section, we demonstrate how easy two prototypes can be developed using our new open-source software framework. We present the implementations of two different recommendation models. We assume that we do not have a dataset we could use for our experiment and for the first implementation, we only want to test our new recommendation model. Therefore, for the first prototype we do not need to develop a new user interface, but can use the user-interface provided by the framework and we need to use the news agency scraper component to populate the database.

5.1 Example Implementation 1: Comment-Centric Comment Recommendation

In the first example implementation[5], we followed an approach from [12] that differs from previous comment recommendation systems. In contrast to comment recommendation systems so far, the suggestions are not based on previous interests and behavior of the user, but on the comments of the discussion. The user select a comment they would like to see different perspectives for and the model retrieves a selection of comments from the database that provide different points of view for the topic of this comment. According to [12], this approach is called comment-centric comment recommendation.

[2] https://pypi.org/project/Comment-Recommendation-Framework/.

[3] https://docs.docker.com/.

[4] https://docs.docker.com/compose/.

[5] https://github.com/hhucn/Example-Implementation-1.

Model. The model consists of a two-step approach. In the first step, it creates a candidate set of comments that it extracts from the database by using the keywords of the article. The candidates are determined using a k-nearest neighbor search using the vector representation of the keywords to find similar articles and taking all comments that appeared under these articles as a candidate set. By this, it reduces the number of possible comments to a manageable amount (Fig. 4 left).

In the next step, it sorts the candidate set based on semantic similarity to the comment the user is interested in (Fig. 4 right). It then takes the top-n comments and send them to the front-end component for rendering.

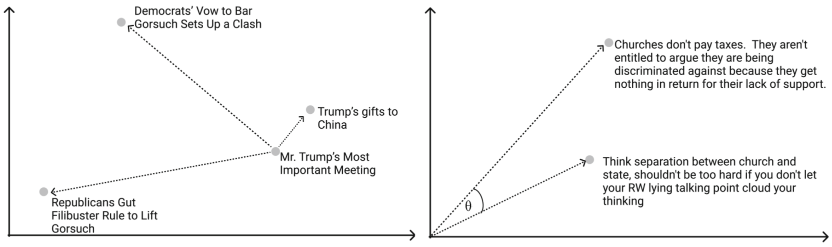
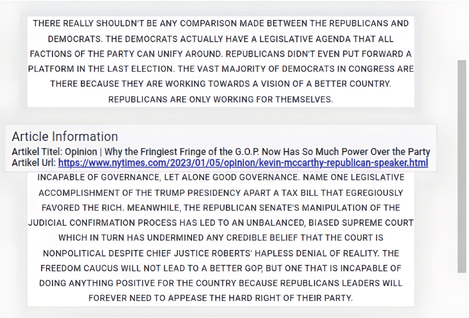

Democrats' Vow to Bar Gorsuch Sets Up a Clash
Trump's gifts to China
Mr. Trump's Most Important Meeting
Republicans Gut Filibuster Rule to Lift Gorsuch
Churches don't pay taxes. They aren't entitled to argue they are being discriminated against because they get nothing in return for their lack of support.
Think separation between church and state, shouldn't be too hard if you don't let your RW lying talking point cloud your thinking
θ

Fig. 4. Left Figure: Find the most similar articles with k-nearest neighbor search. Right figure: Calculate the semantic similarity θ between the comment vectors. Figure from [12]

User-Interface. For the user-interface, we use the ready-to-use built-in UI (Fig. 5) explained in Sect. 4.1. This user-interface renders the suggested comments as a list and provides meta information about the comment while hovering over it, like the article title and URL where the comment has been published.

Implementation Effort. Because we do not have a dataset we could use for our experiment, we have to scrape data from a news agency to test our new model. Therefore, we use the news agency scraper component provided by the package. Here, we only have to write the class that extracts the necessary information from the news agency HTML response. Every news agency has a different HTML structure, hence, we always have to write a class for every news agency we would like to scrape. The rest of the code e.g. to process the data and store them in the database is handled by the framework and has only to be changed if we would like to parse the data in a different way or store other data in the database.

After this, we need to implement the embedding model. Our recommendation model needs vector representations of the article keywords and comments. Therefore, we have to compute these embeddings and store them in the database. Fist, we have to implement the embedding method and then we have to update the script that manages the computation of the embeddings. The script queries



THERE REALLY SHOULDN'T BE ANY COMPARISON MADE BETWEEN THE REPUBLICANS AND DEMOCRATS. THE DEMOCRATS ACTUALLY HAVE A LEGISLATIVE AGENDA THAT ALL FACTIONS OF THE PARTY CAN UNIFY AROUND. REPUBLICANS DIDN'T EVEN PUT FORWARD A PLATFORM IN THE LAST ELECTION. THE VAST MAJORITY OF DEMOCRATS IN CONGRESS ARE THERE BECAUSE THEY ARE WORKING TOWARDS A VISION OF A BETTER COUNTRY. REPUBLICANS ARE ONLY WORKING FOR THEMSELVES.

Article Information

Artikel Titel: Opinion | Why the Fringiest Fringe of the G.O.P. Now Has So Much Power Over the Party

Artikel Url: https://www.nytimes.com/2023/01/05/opinion/kevin-mccarthy-republican-speaker.html

INCAPABLE OF GOVERNANCE, LET ALONE GOOD GOVERNANCE. NAME ONE LEGISLATIVE ACCOMPLISHMENT OF THE TRUMP PRESIDENCY APART A TAX BILL THAT EGREGIOUSLY FAVORED THE RICH. MEANWHILE, THE REPUBLICAN SENATE'S MANIPULATION OF THE JUDICIAL CONFIRMATION PROCESS HAS LED TO AN UNBALANCED, BIASED SUPREME COURT WHICH IN TURN HAS UNDERMINED ANY CREDIBLE BELIEF THAT THE COURT IS NONPOLITICAL DESPITE CHIEF JUSTICE ROBERTS' HAPLESS DENIAL OF REALITY. THE FREEDOM CAUCUS WILL NOT LEAD TO A BETTER GOP, BUT ONE THAT IS INCAPABLE OF DOING ANYTHING POSITIVE FOR THE COUNTRY BECAUSE REPUBLICANS LEADERS WILL FOREVER NEED TO APPEASE THE HARD RIGHT OF THEIR PARTY.

Fig. 5. The recommendations of the model are rendered as a list. If user hover over a certain comment, they are presented with meta information about the comment like the article where it has been posted.

all comments and articles from the database which do not have an embedding, computes the embedding, and then updates the node with it.

In the next step, we need to implement our model. For this, we use the model presented in [12].

At last, we have to consider the user-interface. However, since we use the built-in UI, we do not have any effort here.

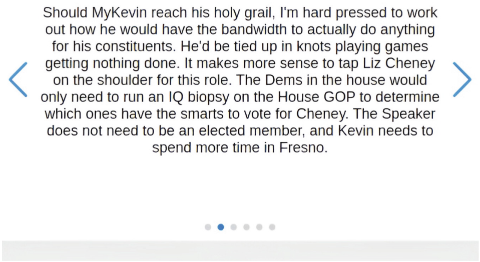
5.2 Example Implementation 2: Most Popular Comment Recommendation

In the second example implementation[6], we follow an approach based on [10]. Here, the model recommends comments based on their popularity, and thus helping the user to get an overview which are the most popular opinions about the topic of the discussion.

Model. The model recommends comments by querying similar articles to the article the user is currently interested in and then creates a candidate set of comments from all these articles. Afterwards, the comments are sorted by the number of up-votes they received to find the most popular comments for the topic of the discussion.

User-Interface. For this example implementation, we replace the built-in user-interface the framework offers with a carousel view to demonstrate how easily the user-interface can be adapted to different circumstances. As we see in Fig. 6, this UI displays one comment at a time and rotates though the list of comments when clicking on the left or right arrow.

[6] https://github.com/hhucn/Example-Implementation-2.



Should MyKevin reach his holy grail, I'm hard pressed to work out how he would have the bandwidth to actually do anything for his constituents. He'd be tied up in knots playing games getting nothing done. It makes more sense to tap Liz Cheney on the shoulder for this role. The Dems in the house would only need to run an IQ biopsy on the House GOP to determine which ones have the smarts to vote for Cheney. The Speaker does not need to be an elected member, and Kevin needs to spend more time in Fresno.

Fig. 6. The recommended comments are displayed one at a time in a carousel view. The user can rotate through the list of comments by clicking the left or right arrow.

Implementation Effort. The implementation effort for this example system is very similar to the example system in Subsect. 5.1. The only differences are that we need to scrape the number of up-votes a comment has received, the user-interface, and implement a different recommendation model.

To scrape the additional data, we need to update the database schema by adding an additional *up_ votes* field to the *Comment* class and by this adding a new property to the comment node in the Neo4J database. Then, we need to update the *process_ item* method in the *pipelines* file to store the new property in the database. The class with the news agency specific methods is nearly the same as in Sect. 5.1 with the exception that we also need to scrape the number of up votes a comment has received.

Next, we update the user-interface. For this, we only have to add a new *carousel component* to the *component* folder of the UI folder where we define how the list of comments should be rendered. Then we have to replace the build in *list* component with the new component in *App.js*.

At last, we need to implement the recommendation model from [10].

5.3 Saved Work

After we have developed the two example implementations and examined the implementation effort, we clearly see the amount of technical code that we did not have to write by using the comment recommendation framework. Let us take a look at the components we did not have to implement.

First, we did not need to implement the complete infrastructure provided by the framework like the user-interface, the API, and the database. Additionally, we use a system that scrapes, processes, and stores the data of news agency sites in the database. At last, all this is already containerized with docker and docker-compose and therefore, can be used on any device that has docker and docker-compose installed.

6 Conclusion

In our work, we introduced an open-source software framework for rapid prototyping of comment recommendation systems to quickly evaluate new algorithms.

With this, we try to relieve researchers of the tedious work of developing prototypes from scratch. While at the same time offering the possibility to evaluate the recommendation model in a realistic setting.

The framework provides a complete modular infrastructure to run a recommendation system in a web application context. This allows to use the system in a laboratory as well as a real-world setting while also reducing the development effort because only prototype specific code has to be written.

Our approach for the development of custom made comment recommendation systems consists of several components which can easily be replaced or updated without major changes to the rest of the system. This allows the researcher to adapt the prototype to new use cases or test e.g. different user-interfaces.

Additionally, we offer a solution for a key point of a comment recommendation system, the data acquisition. A recommendation system can only provide solid suggestions if it has solid data basis. Therefore, we offer two approaches to populate the database for the system. On the one hand, we provide a script to import an existing dataset into the database. On the other hand, we offer a system to scrape arbitrary news agencies sites and store this data in the database.

In future work, we will continue the development of our comment recommendation framework by extending the framework for other kinds of recommendation systems and improving it in general. For this, we will provide more generalized user-interfaces and recommendation model interfaces. Additionally, we will continue to thoroughly test our system in larger field experiments with more sophisticated comment recommendation models and add more features.

References

1. Scrapy (2022). https://scrapy.org/
2. Faridani, S., Bitton, E., Ryokai, K., Goldberg, K.: Opinion space: a scalable tool for browsing online comments. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1175–1184 (2010)
3. Foundation, D.S.: Django (2022). https://www.djangoproject.com
4. Gao, M., Do, H.J., Fu, W.T.: Burst your bubble! an intelligent system for improving awareness of diverse social opinions. In: 23rd International Conference on Intelligent User Interfaces, pp. 371–383 (2018)
5. Hoque, E., Carenini, G.: Convisit: interactive topic modeling for exploring asynchronous online conversations. In: Proceedings of the 20th International Conference on Intelligent User Interfaces, pp. 169–180 (2015)
6. Icons8: Interface, server, database, excavate icons by icons8 (2022). https://icons8.com
7. LaFraniere, S.: Biden administration plans to offer second booster shots to those 50 and up (2022). https://www.nytimes.com/2022/03/25/us/politics/biden-second-booster-shot-older-americans.html
8. Meta Platforms, I.: React framework (2022). https://reactjs.org/
9. Neo4j, I.: Neo4j graph database (2022). https://neo4j.com/
10. Reuver, M., Mattis, N.: Implementing evaluation metrics based on theories of democracy in news comment recommendation (hackathon report). In: Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pp. 134–139 (2021)

11. Shmueli, E., Kagian, A., Koren, Y., Lempel, R.: Care to comment? Recommendations for commenting on news stories. In: Proceedings of the 21st International Conference on World Wide Web, pp. 429–438 (2012)
12. Steimann, J., Feger, M., Mauve, M.: Inspiring heterogeneous perspectives in news media comment sections. In: Yamamoto, S., Mori, H. (eds.) HCII 2022. LNCS, vol. 13305, pp. 118–131. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-06424-1_10
13. Stephens, B.: What if putin didn't miscalculate? (2022). https://www.nytimes.com/2022/03/29/opinion/ukraine-war-putin.html
14. Wang, J., Li, Q., Chen, Y.P.: User comments for news recommendation in social media. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 881–882 (2010)
15. Zhou, M., Shi, R., Xu, Z., He, Y., Zhou, Y., Lan, L.: Design of personalized news comments recommendation system. In: Zhang, C., et al. (eds.) ICDS 2015. LNCS, vol. 9208, pp. 1–5. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24474-7_1

Chapter 5

Understanding the Structural Aspects of Relevance in User Comments

In this chapter, we give an overview about the contributions and impact of our paper Steimann et al. (2024):

Jan Steimann, Markus Brenneis and Martin Mauve.

"Is This Comment More Relevant? Understanding the Structural Aspects of Relevance in Comment Sections"

In: *Proceedings of the 26th International Conference on Human-Computer Interaction*, in the thematic area *Human Interface and the Management of Information*, pages 264–278, Springer International Publishing
Acceptance Rate: ∼55.2%

5.1 Summary

This paper examines whether the structural elements of a user comment, besides its content, can determine its relevance. We aim to explore if an average user finds a comment more relevant when e.g. it includes a justification along with its stance, or even a personal anecdote supporting the justification.

To achieve this, we perform a user study in which participants are presented with several comments on the same topic, each differing by a structural element. They then select the comment they find most relevant.

We devised different relevance characteristics based on prior research, detailed in Section 3 of the paper and divided them into essential and optional elements. We hypothesize that certain characteristics must be present in a comment for it to be perceived as relevant by participants, while others, though unnecessary, enhance perceived relevance.

Contrary to our expectations, the results showed that participants deemed a simple position statement adequate for relevance, without the necessity for further justification. Nonetheless, our hypothesis that the inclusion of structural elements improves the perceived relevance was confirmed. Detailed discussions on these hypotheses can be found in Section 4 of the paper. The outcomes of our user study are presented in Section 5, followed by a discussion in Section 6.

This study aims to develop an intuition of how structural elements impact the perceived importance of user comments. Our objective is to provide an additional perspective to current research methods that evaluate comment relevance based on content and textual characteristics. This insight will be valuable for developing the comment recommendation system discussed in Chapter 6.

5.2 Personal Contribution

Jan Steimann originated the idea of examining whether the relevance of user comments on news agency websites can be determined not only by content but also by structural features. He formulated the hypotheses and executed the user survey, analyzed the outcomes, and assessed which hypotheses were validated and which were not. Markus Brenneis offered advice on whether his evaluation method from Brenneis and Mauve (2020) could be applied to this study, and also gave general feedback on the experiment and paper. The entire concept and experiment were refined through discussions with Martin Mauve, who provided feedback, including on the publication drafts.

5.3 Importance and Impact on this Thesis

This paper establishes the foundation for the recommendation model described in Chapter 6. To construct this model, we require a means to assess the intricate issue of judging the relevance of user comments in news agency comment sections, without considering the broader context of the discussion. Consequently, our objective was to assess the relevance of each comment independently of its content and context by analyzing its structural features. This approach helps break down the complex issue into more manageable parts and enables pre-processing to handle the majority of computationally expensive machine learning tasks prior to obtaining the final recommendations. Previous publications like Rowe (2015) have examined the structural components related to relevance; these form the foundation of our study. However, our work is unique in examining whether the existence or introduction of specific structural elements can increase the perceived relevance of user comments. With this approach, we are equipped to address the challenges outlined in Section 1.1 building our recommendation model in Chapter 6.

Is This Comment More Relevant? Understanding the Structural Aspects of Relevance in Comment Sections

Jan Steimann, Markus Brenneis( ), and Martin Mauve

Heinrich Heine University, Universitätsstraße 1, 40225 Düsseldorf, Germany
Markus.Brenneis@hhu.de

Abstract. What makes a user comment relevant for readers? In this paper, we investigate the structural aspects of user comments as they appear e.g. in the comment sections of news media websites. While other studies already have examined the deliberative quality of user comments across different media, it is not yet well understood how the structural features influence the perceived relevance of a comment. Our goal is to develop an understanding of the influence of various structural aspects like position and different kind of justifications on the perceived relevance of user comments. We approach this question by means of a user study. For this, we ask the participants to decide which of the shown user comments is more relevant in comparison with others. Our study shows that the perceived relevance follows some intuitive rules, e.g. adding more justifications increases the relevance. On the other hand, some results were also surprising, like simple statements without any insights are considered relevant by the users. Our results should support the development of proper metrics and algorithms for comment recommendation systems by providing an additional understanding.

Keywords: Recommender Systems · Comment Recommendation · Information Systems · Humand-Centered Computing

1 Introduction

Discussions in comment sections of news agencies are a popular way to participate in the public debate about a topic. Here, users have the opportunity to state their opinion about the topic of an article and, in the best case, engage in a constructive discussion about their opinion. However, this best case is fairly seldom to see and many discussions become a shouting match or an echo-chamber over time.

For this reason, users need a way to identify relevant contributions which provide new points of view or represent different arguments within the discussion. This task of news article and comment recommendation has been widely examined in the past and various studies have presented interesting and promising approaches.

H. Mori and Y. Asahi (Eds.): HCII 2024, LNCS 14689, pp. 264–278, 2024.
https://doi.org/10.1007/978-3-031-60107-1_19

However in the field of computer science, many approaches for developing new algorithms rely on labeled data-sets where the comment relevance is assessed only on the content level [6,9,11,20]. For example, in [19] the dataset contains scores for how relevant a given comment is on the content level for a specific article and these scores are then used to evaluate the performance of the model.

While these are proven and effective approaches to develop and evaluate new algorithms, assessing the relevance of a comment on the content level is a very difficult task. Each person weighs argumentation differently based on their experience, background, and political orientation. Especially in the context of comment recommendation, there is most likely more than one relevant comment that would be a good recommendation for a given discussion. This makes it even more difficult to develop a robust model that can assess the relevance of comments.

Therefore, we believe that the development of future recommendation algorithms can benefit from a deeper understanding of the structural aspects of what makes a user comment more relevant than another. Studies like [7,13] have already examined the deliberative quality of comments from the structural composition of a comment. [13], for example, has used these structural aspects to compare the deliberative quality of comments across different media and in contiguous discussions. However, it remains to investigate how these features influence the perceived relevance of a single comment in comparison with others.

Does a comment become more relevant if we add more structural features like sources or alternatives? Which features need to be present so that a comment is considered relevant at all? Do some features affect the relevance more than others?

We try to answer these questions through a user survey in which we asked participants to compare comments that state the same opinion but differ in one structural aspect. Then, they had to decide which comment is more relevant than another. For example, all comments state that they like chia seeds very much for breakfast. However, one comment justifies its opinion and we expect that the participants will assess this comment as more relevant than the other comments. With the results of this survey, we want to support the development of future algorithms by providing a deeper understanding of comment relevance backed by human intuition.

In the following section, we take a look at the previous work in this field. In the third section, we outline the different characteristics of a comment we have examined in this study. Afterwards, we explain the methods we used to conduct the survey and to evaluate the results. In the last section before the conclusion, we present the results and interpretation.

2 Related Work

The topic of online commenting has been widely studied from different fields and point of views like computer science, journalism, and computational linguistics [2,12,13,18,21].

Here, users exchange their views on the content of news articles and engage in debates with others. However, the quality and relevance of these comments varies significantly, with some comments being highly informative and insightful, while others being irrelevant or even harmful. Many studies have investigated these aspects and tried to provide an explanation when and how a comment is relevant, a high quality contribution, or harmful in the debate. One example is the tone of the debate. [4] investigates the comment section of news agencies to better understand when and why incivility occurs. The authors found that especially in discussions with a lively debate, incivility was less common. Furthermore, not all incivility is the same for every user. [10] states that people detect and approve incivility differently depending on whether it coincides with their views. This supports our argumentation from Sect. 1 that assessing comment relevance on the content level is a difficult task. Furthermore, some studies [8,13] have even shown that the quality and relevance of a comment differs between different media. In comparison comment sections on news organization sites showed a greater deliberative quality than their online presence on social media like Facebook [13].

This shows that it is worth to consider a wider range of influencing factors in addition to labeled datasets when assessing the quality and relevance of a comment. Future algorithms should consider these influences.

For example, [22] studied the effect of message factors on user comments intention and behavior. On the one hand, they found that certain factors like *aggression*, *incivility*, and *drifting off-topic* increase the likelihood that other users respond in a similar tone, heating up the discussion. On the other hand, deliberative discussion factors like *question* or *additional knowledge* reduce the likelihood of incivility and drifting off-topic in the discussion. This underpins our argumentation that understanding the structural aspects that make one comment more relevant than another is crucial, highlighting the need for better understanding in this area.

Previous recommendation approaches in the field of computer science use content and word-relation features focused approaches for recommending highly relevant comments [1,17,19]. [17] for example uses various textual and discourse relation features like *article length*, *average number of verbs per sentence*, etc. and an annotated dataset where the annotators had to judge whether a comment was relevant for the given article and whether the comment was thoughtful. [19] uses a data-set with different level of relevance for the news and comments and [1] four different categories for the comment article pair like *Relevant*, *Same Entities*, *Same Category*, and *Irrelevant*. All these approaches use an efficient and proved way for developing recommendation algorithms. However, they focus on content and word-relation features and every study uses a different dataset for the problem which makes it difficult to assess how universal there approaches are. Thus, with the results of our survey, we do not want to provide another labeled dataset, but a deeper understanding of the structural aspects of what makes a user comment more relevant than others backed by the human intuition. These aspects are separated from the content level and therefore are not affected by the context of the discussion. We believe that future approaches can develop more robust recommendation models by combining our findings with labeled datasets.

3 Relevance Characteristics for Comments

We now introduce how we model a comment to determine its relevance. We will later compare it against the human intuition of our participants in our survey. The features of this model are based on previous research on this topic [7,13].

3.1 Relevance Characteristics

We now take a look at characteristics of a relevant comment. Our definition is mostly based on [13].

Necessary Characteristics. For a comment to be considered relevant at all, we believe the following necessary characteristics have to be present.

Topic: First and foremost a relevant comment has to address the topic of the discussion. This seems rather obvious, however certain comments only address the topic of the discussion at first glance and then derive to a completely different topic, driving its own agenda. These comments are considered irrelevant by us.

We differentiate two types of topic relevance. First, the *Structuring Topic* which argues about the topic of the discussion e.g. the article addresses the expansion of wind power plants and the comment argues about the unreliability of wind power. Second, we have the *Interactional Topic* which treats a slightly different topic that is however still relevant in the context of the discussion. For example, in a discussion about nuclear power plants, arguments about other forms of power generation are often invoked and provide additional information or perspectives. *Instead of nuclear power plants, we should use wind power plants because...*

Position: Relevant comments have to express their position on a give issue, otherwise they just repeat information already stated in the article or other comments like *Solar power panels are a possibility to generate power*. Though, we want to support the public debate by presenting new point of views to think about to the user and for this the comment has to state a position towards the topic of the discussion.

Justification: Relevant comments have to justify their position. This allows others to judge *the authority of the better argument* [7]. If we only present a selection of positions to the user without any justification, the user would be unable to compare the benefits and downfalls and therefore, it would be much harder to consider the different positions.

Optional Characteristics. In our model the following characteristics are not necessary for a comment to be relevant. However, they can be used to support the justification and by this increase the relevance of the comment. We expect that the different optional characteristics do not contribute equally to the relevance of the comment. In the following, the features are presented in order of their expected relevance for the comment.

Personal Story: Comments can overcome communicative barriers by invoking personal stories. These barriers often arise through a lack of knowledge or about complex issues. A personal story helps to overcome these barriers by warping the complex issue in a more comprehensible form [14]. This is the reason why we expect the personal story to be the most relatable and therefore most relevant additional feature. For example: *I am against the reform of the administration because it will do more harm than good. I work in an administration myself and after the reform the responsibilities were unclear and no one felt responsible for the critical decisions which then have not been made.*

Example: Similar to the personal story, the example uses a concrete situation to make a complex or abstract argumentation more comprehensible. The main difference to the personal story is the missing personal aspect. Nevertheless, we expect the example to be equally relevant as the personal story. For example: *American football is not a more dangerous sport than others. It is only much more prominent in the news. For example, we don't hear about surfers drowning or climbers falling, but these sports can be much more dangerous than football.*

Alternative: Comments that attempt to provide a solution to the problem at hand by offering an alternative are considered more relevant than comments that provide additional sources because they offer another point of view on the topic. Nevertheless, we expect this characteristic to be not as relevant as the example and personal story because the alternative solution is not as relatable as a personal story or example. For example *I reject the conclusion of the article that we need to replace all old houses with more energy-effective homes. The demolition of the old and construction of a new house is far more energy intensive that just modernizing old houses that are still in good shape.*

Source: By using and referencing additional sources comments provide users with the opportunity to verify the quality and validity of their justification and by this making the comment more relevant. Nevertheless, we expect this characteristic to be the least relevant feature because sources refer to external material which the users need to consider while the other characteristics provide all their information at hand.

Remember that not all of these characteristics are necessary so that a comment is considered relevant. However, at least a comment needs to address the *topic*, *position*, and *justification* dimension to be relevant in our context and therefore provide new point of views on the topic of the discussion and by this fuel the public debate. The remaining dimensions support the *justification* and make a comment more relevant.

Nevertheless, we did not incorporate all quality dimensions which come to mind in terms of comment relevance like *interaction* or *asking questions*, mentioned in [13]. This is because paper like [13] focus on comments only in the context of contiguous discussions where the deliberation occurs through the interaction of participants by referring or responding to each other. However, in the context of comment recommendation such comments that refer to others are not always that relevant as in the context of the contiguous discussion they have been published. Yet, the comments we are interested in, have to be relevant in connection with the topic of the discussion while at the same time have to be understood isolated. We want to incorporate comments from various discussions and news agencies to provide multiple perspectives on the discussion. Therefore, if the comments are only understood in the context of their original discussion, they are not relevant for us.

3.2 Example

An example for a relevant comment[1] in a discussion about solar power panels as a renewable energy source could look like this:

> Everyone with their own house should install solar plant panels to power their house. **(Position and Topic)**
> You get free power all the year and during the day you power the neighborhood with clean energy helping the environment. **(Justification)**
> For example, you could install them on your garage and power your electric car with it. **(Example)**
> I had solar plant panels installed years ago, with an estimated payback of 15-17 years. However, we also acquired two electric cars and charge them at home. The savings in gasoline alone took the solar system payback down to under 3 years. **(Personal Story)**
> Additionally, as explained here: *www.somewebsite.com*, sun and wind power are not "too unreliable". The sun is only unreliable if we look from a local point of view. It is always shining on half of the earth at any given time. If the world were connected into a single electrical distribution network, the sun would be perfectly reliable all the time.**(Sources)**
> And even if you don't own a house and live in a rental apartment, you can rent portable solar panels to reduce your energy consumption from your power supplier. **(Alternative)**

4 Methods

In this section, we explain how we developed our hypotheses to check if our characteristics for a relevant comment match the perception of humans. We

[1] The example is based on these comments: https://nyti.ms/327zNHC#permid=107994755, https://nyti.ms/3tYGpVa#permid=107994566.

developed a questionnaire with scenarios for every hypothesis, and participants of the survey were asked to compare the displayed comments and decide which of the given comments was more relevant or most relevant. In order to achieve a higher information value for our results and to avoid any biases that might occur towards certain topics, we have used different topics with pro and contra position. For every hypothesis one topic with the pro and contra position was randomly chosen. Every participant had to answer all 14 hypothesis which results in 14 topics and 28 comments for every participant.

We selected two more neutral topics *Chia Seeds* and *Boosting the Immune System* and two more controversial topics *Solar Power Panels* and *American Football*. The reason for this is that we want to know if our relevance characteristics apply for neutral topics as well as more controversial topics where many users have a strong opinion. Even if some users assess some comments based on their personal opinion, we can still get valuable data for our hypotheses. For example, if a user is heavily biased against solar power panels and is not open to any arguments in favor of them, we have always an opposing argument for the same topic which they will most likely assess not biased.

The comments we use for our study are all based on comments we extracted from real conservation's in comment sections and have been transformed and reshaped into building blocks so that we can present the same comment that differs in exactly one structural feature to the participants. This way, we can clearly associate the results with a certain characteristic. If we would use the comments as they appeared in the comment sections, it would be difficult to present the same comment with exactly one different structural feature because these comments use them intertwined with each other. For example, a comment might combine its position towards the topic with its justification. Therefore, we would need to find a different comment with exactly the same position that differs exactly in this specific structural characteristic. However, then we could not differentiate if the results are based on the additional characteristic or the different writing style of the two comments.

Before the users started with the survey, we presented them with a definition what makes a comment more relevant than another comment. For this, the comment has to *provide a more thoughtful or elaborate perspective in the discussion than the other comment*. This definition is the essence of our definition for a relevant comment from Sect. 3. We believe that a comment becomes more relevant the more it elaborates its position on the given topic with justifications and additional characteristics like personal stories, sources, examples, or alternatives. At the same time the definition which we presented to the participants of our survey must not be too detailed to prevent us from influencing the participants to much.

The complete list of hypotheses is in Table 1. At first, we were interested in the minimal requirements for a comment to be considered relevant. Afterwards, we examined how a comment becomes more relevant, either by adding more justifications or by supporting the justification with additional characteristics. Then we asked ourselves if there exists an order between the different characteristics that support the justification. For example, do users tend to consider a personal story more relevant than supporting the argumentation with additional sources? Additionally, we asked our self if a comment becomes more relevant if we provide more supporting characteristics of the same type or if these are of lower importance the user.

As an example, we now present how Hypothesis 4 *(A comment becomes more relevant if it backs up its justification with sources, a personal story, an alternative, or an example)* has been developed and transformed in a questionnaire scenario.

We were interested whether adding a supporting characteristic like a personal story or sources to an existing comment, adds more relevance to the comment. Therefore we can assume in future metrics or recommendation models that if a comment contains an additional characteristic, it can be considered more relevant than the same comment without.

From our hypothesis, we constructed the following comments where the users should answer different questions like *Is Alice's comment more relevant than Bob's?* to determine which comment is more relevant:

Alice writes:
Everyone with their own house should install solar power panels.

Bob writes:
Everyone with their own house should install solar power panels because they provide you with free energy.

Charlie writes:
Everyone with their own house should install solar power panels because they provide you with free energy.
However, even if you don't own a house you can rent solar panels and install them on your balcony to reduce your electricity bill.

We expect that *Charlie's* comment is the most relevant one because he additionally provides an *alternative solution* for the topic at hand. We added the other comment to prevent biases in the answers.

We created the questions and scenarios for the remaining hypotheses in the same way.

Table 1. Our hypothesis about the structural aspects of relevance in comment sections.

#	Hypothesis
H1	*A comment needs to address the topic, take a position towards the topic, and provide a justification to be relevant.*
H2	*A comment with an interactional topic will be considered relevant, but not as relevant as a comment with structuring topic.*
H3	*Providing more than one justification makes a comment more relevant than one justification.*
H4	*Backing up the justification with additional characteristics like sources, personal stories, alternatives, or example makes the comment more relevant.*
H5	*Providing more than one additional characteristic makes the comment more relevant.*
H6	*Using an alternative instead of a source makes the comment more relevant.*
H7	*Using a personal story instead of an alternative makes the comment more relevant.*
H8	*Using an example or a personal story makes the comment equally relevant.*
H9	*A comment with an alternative is more relevant than the same comment with sources, but less relevant than a comment with a personal story or an example.*
H10	*Providing more than one source makes the comment more relevant.*
H11	*Providing more than one personal story makes the comment more relevant.*
H12	*Providing more than one alternative makes the comment more relevant*
H13	*Providing more than one example makes the comment more relevant.*
H14	*A comment with an example is equally relevant as a comment with a personal story. But example and personal story are more relevant than an alternative.*

5 Results

In the following section, we present our results of the survey. For the evaluation, we used the method presented in [3] where we report the p-value for the null hypothesis H_0 that *Our expected answer is not the most frequently (relative frequency) given answer.*[2]

We conducted the survey with Amazon Mechanical Turk. We are well aware that MTurk needs to be treated with caution due to bots and randomly clicked answers by the workers. For this reason we used several carefully developed control questions to ensure that only qualified worker participate in our survey. Of the 82 participants of our survey, only 36 answered enough control questions correctly to meet our quality standards. These 36 workers had an median age of 30–39 years, and 15 men and 21 women participated.

Furthermore, we should note that our sample of worker is not representative for the US population, and even less world wide and the sample size is insufficient to assert the universality and representatives of our findings. Consequently, we cannot completely generalize our results and have to use them with caution. Nevertheless, the insights we gathered offer valuable contributions to enhance the understanding and develop an intuition regarding the structural aspects of relevance for future studies and recommendation algorithms.

However, our results confirmed our most important hypotheses and we also made some surprising findings. In the next section, we will talk about the hypotheses that were confirmed and afterwards, we discuss the remaining results. Additionally, a figure of our results can be found in our data repository[3].

Before proceeding, as we explained in Sect. 4, we ensured that there is no bias due to the chosen topics. To validate the absence of such a bias, we computed the average percentage of expected answer for all hypotheses and questions related both neutral and controversial topics.

Our analysis revealed no significant differences between the results obtained from the controversial and neutral topics. The average percentage of expected answers for neutral topics was 52%, while for controversial topics, it was 50%. These results underscore the consistency of our results and reaffirm its suitability for drawing unbiased conclusions.

5.1 Results that Confirmed Our Hypotheses

In hypothesis H2, we wanted to investigate how user assess the relevance of a comment that states a position with an interactional topic in comparison with a comment that provides a structuring topic. An interactional topic addresses a slightly different topic, which is still relevant in the context of the discussion, while a structuring topic matches the subject of the discussion. For H2, the expected answer was given by 56% of the participants with a p-value of 0.02. This strongly supports our hypothesis that the interactional topic is considered more relevant than a comment that just states a position towards the topic and

[2] We used an intersection-union test [16] with a one-tailed z-test on the variance of the two multinomial proportions [5,15], i.e. H_0 is that the differences of the relative answer frequencies between the expected answer and the other answers is not greater than 0.

[3] https://github.com/hhucn/Comment-Relevance-Survey-Results.

slightly less relevant than a comment with a topic that is a little bit more on point.

An essential aspect for us was to understand if we can make a comment more relevant by additively adding more structural aspects to the comment or if the relevance is perceived in a different way that does not allow an additive consideration.

To check this property we started with hypothesis H3, where we investigated if a comment becomes more relevant if we add more than one justification to the comment. The expected answer was given by 61% of the participants ($p<0.001$), which strongly suggests that our hypothesis is true.

Afterwards we tried to understand what makes a justification more relevant. For this, we have explored if a comment becomes more relevant if the justification is supported by additional characteristics like a personal story, alternative solution, sources, or an example (H4), or if the users do not perceive these as separate building blocks and focus only on the content of the comment. 61% provided us with the anticipated answer ($p = 0.006$), which confirms our hypothesis that a comment becomes more relevant if we support the justification with additional characteristics.

The logical step is to investigate if the idea of hypothesis 3 applies also to the justification and we can make it more relevant by additively adding more characteristics to the justification (H5). This was confirmed (58%, $p = 0.001$).

Following, we have investigated the additional characteristics in more detail. We wanted to find out if the users recognise more than one additional characteristic of the same type as separate entities or just as a larger single entity. This would have different consequences for the perceived relevance of the comment because e.g. providing different examples for a justification gives a comment more credibility than just a single example. First, we provided more than one additional source (H10) and received the expected answer by 61% ($p < 0.001$) which confirms our hypotheses with a high confidence. Next, we investigated if providing more than one personal story makes a comment more relevant (H11), which was confirmed (55%, p-value = 0.012). Afterwards, we did the same for the alternative (H12) with 52% providing the expected answer (p-value = 0.034).

5.2 Unexpected Results

After we have discussed the results that confirmed our hypotheses, we now examine the results that surprised us or where further investigations are needed.

First and foremost in our survey, we wanted to understand if there exists a structural threshold from which a comment can be considered relevant. Our hypothesis was that a comment needs to contain certain features to be considered relevant (H1). First, the comment needs to address the topic of the article and state a position towards it and second it needs to justify its position. We thought that especially the justification is crucial for a comment to be considered relevant because it allows users to judge the different argumentations. However, the expected answer was given only by 39%; whereas 58% considered the comment that just states a position towards the topic relevant. This indicates that the users set the threshold for a relevant comment much lower than we expected. Thus, it is still necessary to examine in more detail at what point users classify a comment as relevant at all.

Subsequently, we investigated the additional characteristics for the justification. As we explained in Sect. 4, we wanted to understand if there exists an order of relevance for the different additional characteristics like personal story, example, alternative solution, and sources. For example, is a personal story more relevant than sources because the personal story helps the users to comprehend a complex issue by wrapping it in a relatable story? However, the results indicate that the users did not perceive some additional characteristics more relevant than others. In H6, we investigated if a comment with an alternative is more relevant than a comment with sources. Here, only 31% selected the alternative as the most relevant comment. In H7, we investigated if a personal story is more relevant than an alternative (expected answer by 43% for the personal story). In H8, we expected that a personal story and an example are equally relevant. However, only 38% gave the expected answer. In H9, we directly compared the personal story, alternative, and sources, and supposed that the personal story is the most relevant comment, the alternative the second most relevant, and the sources the least relevant of the three comment. However, only 30% gave the expected answer. In the last hypothesis (H14), we put the personal story in direct comparison with an example where we assumed that they are equally relevant because they only differ in the personal component (expected answer by 21%). However, the results indicate that the comment with the personal story is perceived as the most relevant (51%) which is contrary to the results of hypothesis H9 where the comment with the personal story was not the most relevant one. Therefore, a more detailed investigation here is needed.

The last hypothesis which was not confirmed was H13. Here, we investigated if a comment becomes more relevant if we provide more than one example. Here, only 47% considered the comment with two examples more relevant, though, 47% indicate that the comment was indeed perceived slightly more relevant.

6 Discussion

After examining the results of the survey, many hypotheses were confirmed. First and foremost, it could be confirmed that we can increase the relevance of a comment by additively adding more structural features like sources, personal story, etc. independent of the content of the comment. We also gained a deeper understanding how additional characteristics like sources, personal story, example, or alternative solution influence the perceived relevance of a comment. For example, we can confirm that the user perceive a comment as more relevant just by adding more than one personal story or source.

However, some results were unexpected and need further investigation. The most unexpected result is that users perceive the threshold for when a comment is relevant much lower than expected. Many user considered a comment as relevant even if it only addresses the topic of the article. We assumed that a comment needs to state a position towards the topic and justify it to be considered relevant. This is a very interesting result and needs to be investigated in more detail to find out where the user draw the line for when a comment is not considered relevant any more or whether this is to subjective to have a threshold here.

Most of the other not confirmed hypotheses focus much more on the details of the structural characteristics like if a personal story is more relevant than sources. Here, we expected that e.g. a personal story is considered more relevant than sources because the personal story wraps a complex issue in a comprehensible format that can be processed more easily.

Still, most of our key hypotheses were confirmed with a high significance. Our main point was to see if we can increase the relevance of a comment by adding more structural features like one or more justifications or making a justification with e.g. a personal story more compelling.

As explained in Sect. 5, we are well are that our results are not completely generalizable due to our limited worker sample and artificial setting. However, we still gathered some valuable insights which help to develop a better intuition regarding the structural aspects of comment relevance.

7 Conclusion and Future Work

In this study, we have conducted a survey with human participants about the structural aspects of comment relevance in news agency comment sections. Our results help to deepen the understanding of comment relevance and provide an additional angle for the development of more sophisticated machine learning and recommendation models. Most of our important hypotheses were confirmed and we showed that we can increase the relevance of a comment additively by adding more structural aspects like justifications. We were also able to improve the justification of a comment by adding additional characteristics like personal story, alternative, example, or sources.

However, some of our hypotheses were not confirmed, e.g. the threshold when a comment is considered relevant is much lower as we expected. Second, we tried to understand if there is an order between how much relevance the additional characteristics could add to the relevance of a comment. For example, we assumed that a personal story makes a comment more relevant than sources.

Nevertheless, this does not automatically mean that there is no just order, but more research is needed here. In future work, we will refine our results and investigate the hypotheses that were not confirmed. Here, it will be of great interest to understand if there exists a threshold for a comment to be considered relevant or if this is too subjective. Another interesting aspect will be to investigate the additional characteristics in more detail.

Yet, with the results that confirmed our hypotheses, we can use these new insights in comment relevance for the development of future comment recommendation algorithms that consider both content and structural aspects for the recommendation.

References

1. Alshehri, J., Stanojevic, M., Dragut, E., Obradovic, Z.: Stay on topic, please: aligning user comments to the content of a news article. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021. LNCS, vol. 12656, pp. 3–17. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72113-8_1
2. Ben-David, A., Soffer, O.: User comments across platforms and journalistic genres. Inform., Commun. Society **22**(12), 1810–1829 (2019)
3. Brenneis, M., Mauve, M.: Do I Argue Like Them? A Human Baseline for Comparing Attitudes in Argumentations. In: Fazzinga, B., Furfaro, F., Parisi, F. (eds.) Proceedings of the Workshop on Advances in Argumentation in Artificial Intelligence 2020, pp. 1–15. No. 2777 in CEUR Workshop Proceedings, Aachen (Nov 2020). http://ceur-ws.org/Vol-2777/paper21.pdf
4. Coe, K., Kenski, K., Rains, S.A.: Online and uncivil? patterns and determinants of incivility in newspaper website comments. J. Commun. **64**(4), 658–679 (2014)
5. Franklin, C.H.: The 'margin of error' for differences in polls (2007). https://abcnews.go.com/images/PollingUnit/MOEFranklin.pdf
6. Gao, M., Do, H.J., Fu, W.T.: Burst your bubble! an intelligent system for improving awareness of diverse social opinions. In: 23rd International Conference on Intelligent User Interfaces, pp. 371–383 (2018)
7. Habermas, J.: The structural transformation of the public sphere: An inquiry into a category of bourgeois society. MIT press (1991)
8. Hille, S., Bakker, P.: Engaging the social news user: Comments on news sites and Facebook. Journal. Pract. **8**(5), 563–572 (2014)
9. Hoque, E., Carenini, G.: Convisit: Interactive topic modeling for exploring asynchronous online conversations. In: Proceedings of the 20th International Conference on Intelligent User Interfaces, pp. 169–180 (2015)
10. Muddiman, A., Stroud, N.J.: News values, cognitive biases, and partisan incivility in comment sections. J. Commun. **67**(4), 586–609 (2017)
11. Mullick, A., Ghosh, S., Dutt, R., Ghosh, A., Chakraborty, A.: Public sphere 2.0: targeted commenting in online news media. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II, pp. 180–187. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-15719-7_23
12. Reimer, J., Häring, M., Loosen, W., Maalej, W., Merten, L.: Content analyses of user comments in journalism: a systematic literature review spanning communication studies and computer science. Digital Journalism, pp. 1–25 (2021)
13. Rowe, I.: Deliberation 2.0: Comparing the deliberative quality of online news user comments across platforms. J. Broadcast. Electron. Media **59**(4), 539–555 (2015)
14. Ryfe, D.M.: Narrative and deliberation in small group forums. J. Appl. Commun. Res. **34**(1), 72–93 (2006)
15. Scott, A.J., Seber, G.A.: Difference of proportions from the same survey. Am. Stat. **37**(4a), 319–320 (1983)
16. Silvapulle, M.J., Sen, P.K.: Constrained statistical inference: Order, inequality, and shape constraints. John Wiley & Sons (2011)
17. Swapna, G., Jiang, J.: Finding thoughtful comments from social media. ACL (2012)
18. Weber, P.: Discussions in the comments section: Factors influencing participation and interactivity in online newspapers' reader comments. New media society **16**(6), 941–957 (2014)

19. Wei, H., Zheng, W., Xiao, Y., Dong, C.: News-comment relevance classification algorithm based on feature extraction. In: 2021 International Conference on Big Data Analysis and Computer Science (BDACS), pp. 149–152. IEEE (2021)
20. Zhou, M., Shi, R., Xu, Z., He, Y., Zhou, Y., Lan, L.: Design of personalized news comments recommendation system. In: Zhang, C., et al. (eds.) ICDS 2015. LNCS, vol. 9208, pp. 1–5. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24474-7_1
21. Ziegele, M., Johnen, M., Bickler, A., Jakob, I., Setzer, T., Schnauber, A.: Male, hale, comments? factors influencing the activity of commenting users on online news websites. Stud. Commun. Media **2**(1), 110–114 (2013)
22. Ziegele, M., Weber, M., Quiring, O., Breiner, T.: The dynamics of online news discussions: effects of news articles and reader comments on users' involvement, willingness to participate, and the civility of their contributions. Inform., Commun. Society **21**(10), 1419–1435 (2018)

Chapter 6

Bringing Everything Together – Developing A Comment Recommendation Model

In this chapter, we give an overview about the contributions and impact of our paper Steimann and Mauve (2025):

Jan Steimann and Martin Mauve.

"What Do Other People Think About This? Recommending Relevant and Divers User Comments in Comment Sections"

Manuscript submitted for publication

6.1 Summary

This paper serves as the cornerstone of this dissertation, building upon all preceding publications within it. We present a model to tackle the novel recommendation task as detailed in Steimann et al. (2022). The model generates recommendations that are not only pertinent to the user comment but also offer a broad range of perspectives.

To achieve this, the recommendation model advances the basic model outlined in Chapter 3, which uses semantic similarity to derive suitable recommendations. It builds on this by integrating structural features from relevant user comments discussed in Section 5 and is implemented using the open-source framework described in Section 4.

The initial section of the paper provides motivation for the necessity of such a recommendation system. Following this, various publications that explore similar research methods are examined, and it is clarified how this research approach fits into the current body of research and how it will enhance it.

Section 3 of the paper delves into the model's specifics. Initially, the purpose and range of application are outlined to clarify its use-case scenarios, alongside a discussion on its constraints and ethical implications. Given these considerations, it is emphasized that this recommendation model serves as a provisional phase in the journey towards creating a publicly accessible model.

Subsequently, the core concept is explained regarding how the model derives its recommendations through the use of a recommendation score. The following section then delves into a thorough explanation of how this recommendation score is determined and the specifics of the model's implementation. In this section, we learn that the model expands on the two-step method outlined in Steimann et al. (2022). This approach is aimed at minimizing the potential comment candidate set to improve retrieval speed. The model leverages the specified keywords from the article where the user's comment is located to discover topics and articles aligned with the article's theme. Afterwards, sub-discussions from these articles are identified that fit the user comment thematically by utilizing short text clustering. Through this clustering procedure, thematically relevant comments to the user comment are identified. Finally, the relevance and position of the comments are evaluated to assess their suitability as a recommendation. This process incorporates structural aspects from Steimann et al. (2024) alongside other machine learning approaches such as stance detection and emotion classification, among others.

Subsequently, we evaluate the recommendation model against a random baseline. The findings show that the model surpasses the random baseline but also point out challenges needing further investigation. The paper concludes by summarizing and offering a future work outlook.

6.2 Personal Contribution

As discussed in Section 5.3, Jan Steimann conceived the idea of utilizing structural elements to evaluate the relevance of user comments. He extended this concept into a recommendation model by incorporating various viewpoints through techniques such as stance detection, sentiment and emotion analysis, and the consideration of the news outlet hosting the comments. He implemented this model with the framework described in Section 4 and created the web application for testing the recommendation system. He also carried out the user evaluation. Furthermore, Jan Steimann has written the paper. The formulation of the ideas for the recommendation model and user study was improved through discussions with Martin Mauve, who also reviewed the draft of the manuscript.

6.3 Importance and Impact on this Thesis

The model introduced in this chapter represents the culmination of the previous publications contained in this thesis. To the best of our knowledge, it is the first paper to present a selection of relevant user comments offering a range of viewpoints on a given comment in comment sections of news agencies. Alongside the initial publication examined in Chapter 3, it proposes the first model addressing the task described in Steimann et al. (2022). This task aims to enhance current research on comment recommendation by delivering diverse perspectives to users engaged in public discourse. Nonetheless, it is crucial to recognize that this model is only a temporary stage in the journey toward creating a functional public-facing model, as various challenges, such as preventing the spread of misinformation and the restriction to a limited range of topics, still need to be addressed.

What Do Other People Think About This? Recommending Relevant and Diverse User Comments in Comment Sections

Jan Steimann[*]
jan.steimann@hhu.de
Heinrich-Heine-University
Düsseldorf, NRW, Germany

Martin Mauve
Heinrich-Heine-University
Düsseldorf, NRW, Germany

Abstract

Online news platforms' comment sections allow individuals to engage in public discussions. People, however, often visit the website of online news agencies that conform to their political beliefs. Consequently, this results in limited encounters with opinions differing from their own, restricting exposure to diverse viewpoints. However, encountering a variety of perspectives is essential for forming well-rounded opinions. In response, we propose a novel comment recommendation model that offers a selection of diverse viewpoints related to a given user comment. Our aim is to address the question "What Do Other People Think About This?" by exposing users to viewpoints they might not normally encounter.

CCS Concepts

• **Computing methodologies** → *Information extraction*; *Knowledge representation and reasoning*; *Search methodologies*; • **Applied computing** → Sociology.

Keywords

Comment Recommendation, Natural Language Processing, Recommendation Model

ACM Reference Format:
Jan Steimann and Martin Mauve. 2018. What Do Other People Think About This? Recommending Relevant and Diverse User Comments in Comment Sections. *ACM Trans. Graph.* 37, 4, Article 111 (August 2018), 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Over the years, the way people interact with news and articles has evolved. Historically, the majority of people passively absorbed news content, with only a small number taking the initiative to compose a letter to the editor to express their views regarding the article.

With the advent of the Internet, this dynamic has shifted. People are now able to effortlessly share their opinions about the article's subject in the comment section. This shift converts individuals from passive consumers into active contributors, enabling them to participate in public discourse. They can voice their viewpoints and also explore new perspectives that they might not have encountered otherwise.

Yet, as shown by [Flaxman et al. 2016], many individuals act as they would in conventional offline environments, often going straight to the homepage of their preferred news source, which restricts their exposure to diverse perspectives. Even when individuals do not engage in public discourse and merely read, they encounter a limited viewpoint on the subject.

Additionally, this novel method of communication introduces not only benefits to public discussions but also fresh challenges. On one hand, allowing everyone to engage in public debates can quickly make discussions overwhelming because of the influx of comments that may be generated. It is frequently the case that an article attracts several hundred comments, or even a few thousand, particularly if it pertains to a controversial subject. In this scenario, recommendation systems can assist users in locating the most relevant comments for them or the discussion.

On the other hand, just recommending the most relevant comments of a discussion is hardly enough to foster the public discourse. As research suggests, the effective functioning of a democratic society depends greatly on its citizens having access to a broad spectrum of news and perspectives [Joris et al. 2020].

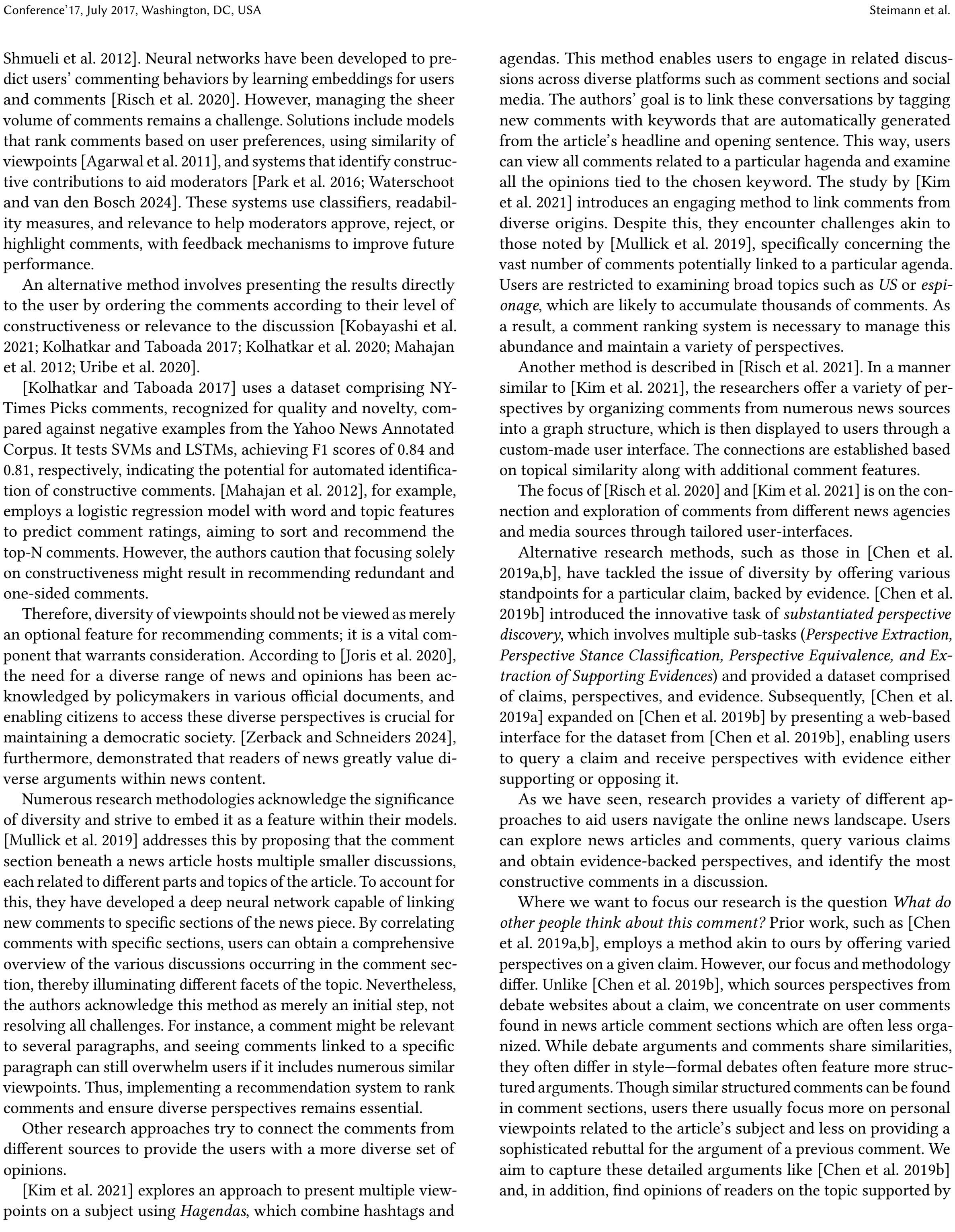
Our objective is to advance the field of comment recommendation by assisting users in forming opinions using a wider range of viewpoints on topics they care about. Unlike conventional recommendation systems that align with a user's interests or behavior, our goal is to present users with a diverse set of perspectives on discussion topics, thereby facilitating their own opinion development. This study introduces our recommendation model, which suggests comments drawn from various discussions and distinct communities, depending on the comment currently being engaged with by the user. The central question we address is: *What Do Other People Think About This?* This research aims to supplement the existing work, which is elaborated upon in Section 2 on related work.

2 Related Work

Several research paper on content recommendation for news agencies have highlighted the role of personalized systems in helping users manage the large volume of online content. While recommending news articles tailored to user interests is a well-known application [Li et al. 2010; Liu et al. 2024], recommendation systems can also enhance user engagement in public debates, allowing users to find discussions they are interested in [Risch et al. 2020;

Authors' Contact Information: Jan Steimann, jan.steimann@hhu.de, Heinrich-Heine-University, Düsseldorf, NRW, Germany; Martin Mauve, Heinrich-Heine-University, Düsseldorf, NRW, Germany.

ACM 1557-7368/2018/8-ART111
https://doi.org/XXXXXXX.XXXXXXX



Shmueli et al. 2012]. Neural networks have been developed to predict users' commenting behaviors by learning embeddings for users and comments [Risch et al. 2020]. However, managing the sheer volume of comments remains a challenge. Solutions include models that rank comments based on user preferences, using similarity of viewpoints [Agarwal et al. 2011], and systems that identify constructive contributions to aid moderators [Park et al. 2016; Waterschoot and van den Bosch 2024]. These systems use classifiers, readability measures, and relevance to help moderators approve, reject, or highlight comments, with feedback mechanisms to improve future performance.

An alternative method involves presenting the results directly to the user by ordering the comments according to their level of constructiveness or relevance to the discussion [Kobayashi et al. 2021; Kolhatkar and Taboada 2017; Kolhatkar et al. 2020; Mahajan et al. 2012; Uribe et al. 2020].

[Kolhatkar and Taboada 2017] uses a dataset comprising NYTimes Picks comments, recognized for quality and novelty, compared against negative examples from the Yahoo News Annotated Corpus. It tests SVMs and LSTMs, achieving F1 scores of 0.84 and 0.81, respectively, indicating the potential for automated identification of constructive comments. [Mahajan et al. 2012], for example, employs a logistic regression model with word and topic features to predict comment ratings, aiming to sort and recommend the top-N comments. However, the authors caution that focusing solely on constructiveness might result in recommending redundant and one-sided comments.

Therefore, diversity of viewpoints should not be viewed as merely an optional feature for recommending comments; it is a vital component that warrants consideration. According to [Joris et al. 2020], the need for a diverse range of news and opinions has been acknowledged by policymakers in various official documents, and enabling citizens to access these diverse perspectives is crucial for maintaining a democratic society. [Zerback and Schneiders 2024], furthermore, demonstrated that readers of news greatly value diverse arguments within news content.

Numerous research methodologies acknowledge the significance of diversity and strive to embed it as a feature within their models. [Mullick et al. 2019] addresses this by proposing that the comment section beneath a news article hosts multiple smaller discussions, each related to different parts and topics of the article. To account for this, they have developed a deep neural network capable of linking new comments to specific sections of the news piece. By correlating comments with specific sections, users can obtain a comprehensive overview of the various discussions occurring in the comment section, thereby illuminating different facets of the topic. Nevertheless, the authors acknowledge this method as merely an initial step, not resolving all challenges. For instance, a comment might be relevant to several paragraphs, and seeing comments linked to a specific paragraph can still overwhelm users if it includes numerous similar viewpoints. Thus, implementing a recommendation system to rank comments and ensure diverse perspectives remains essential.

Other research approaches try to connect the comments from different sources to provide the users with a more diverse set of opinions.

[Kim et al. 2021] explores an approach to present multiple viewpoints on a subject using *Hagendas*, which combine hashtags and agendas. This method enables users to engage in related discussions across diverse platforms such as comment sections and social media. The authors' goal is to link these conversations by tagging new comments with keywords that are automatically generated from the article's headline and opening sentence. This way, users can view all comments related to a particular hagenda and examine all the opinions tied to the chosen keyword. The study by [Kim et al. 2021] introduces an engaging method to link comments from diverse origins. Despite this, they encounter challenges akin to those noted by [Mullick et al. 2019], specifically concerning the vast number of comments potentially linked to a particular agenda. Users are restricted to examining broad topics such as *US* or *espionage*, which are likely to accumulate thousands of comments. As a result, a comment ranking system is necessary to manage this abundance and maintain a variety of perspectives.

Another method is described in [Risch et al. 2021]. In a manner similar to [Kim et al. 2021], the researchers offer a variety of perspectives by organizing comments from numerous news sources into a graph structure, which is then displayed to users through a custom-made user interface. The connections are established based on topical similarity along with additional comment features.

The focus of [Risch et al. 2020] and [Kim et al. 2021] is on the connection and exploration of comments from different news agencies and media sources through tailored user-interfaces.

Alternative research methods, such as those in [Chen et al. 2019a,b], have tackled the issue of diversity by offering various standpoints for a particular claim, backed by evidence. [Chen et al. 2019b] introduced the innovative task of *substantiated perspective discovery*, which involves multiple sub-tasks (*Perspective Extraction, Perspective Stance Classification, Perspective Equivalence, and Extraction of Supporting Evidences*) and provided a dataset comprised of claims, perspectives, and evidence. Subsequently, [Chen et al. 2019a] expanded on [Chen et al. 2019b] by presenting a web-based interface for the dataset from [Chen et al. 2019b], enabling users to query a claim and receive perspectives with evidence either supporting or opposing it.

As we have seen, research provides a variety of different approaches to aid users navigate the online news landscape. Users can explore news articles and comments, query various claims and obtain evidence-backed perspectives, and identify the most constructive comments in a discussion.

Where we want to focus our research is the question *What do other people think about this comment?* Prior work, such as [Chen et al. 2019a,b], employs a method akin to ours by offering varied perspectives on a given claim. However, our focus and methodology differ. Unlike [Chen et al. 2019b], which sources perspectives from debate websites about a claim, we concentrate on user comments found in news article comment sections which are often less organized. While debate arguments and comments share similarities, they often differ in style—formal debates often feature more structured arguments. Though similar structured comments can be found in comment sections, users there usually focus more on personal viewpoints related to the article's subject and less on providing a sophisticated rebuttal for the argument of a previous comment. We aim to capture these detailed arguments like [Chen et al. 2019b] and, in addition, find opinions of readers on the topic supported by a personal view. In contrast to us [Chen et al. 2019b] emphasizes more on extracting well-supported arguments for specific claims.

Work like [Kim et al. 2021] and [Risch et al. 2021] which try to connect comments from different sources and by this incorporating different points of views as well, adopt also a comparable method to our approach. Nevertheless, they emphasize the overview and exploration aspects more than we do. Our aim is to offer users an in-depth perspective in the discussion with different viewpoints for the given comment at hand and therefore provide a complement to this exploration research. Our aim is to help users form informed opinions by offering various viewpoints from diverse communities regarding the comment that captivates the user.

3 Recommendation Model

In this section, we explore the functionality of the recommendation model[1]. Initially, we provide a general summary of its aim and limitations, followed by an in-depth examination of its specific components.

3.1 Aim/Scope of the model

Our model is intended for usage in the comment sections of news articles where users engage in discussions about the article's content. Its purpose is to compile an overview of the diverse viewpoints about an argument or opinion presented by another commenter. When readers encounter a comment that piques their interest—referred to hereafter as **user comment**—they might be curious about alternative views regarding it. This is where our model is beneficial. It examines the user's selected comment alongside the article's keywords to pull relevant comments from a database, which includes inputs from various articles and news sources addressing the same topic. The model provides an array of comments that either support or challenge the position of the user comment, delivering a spectrum of perspectives for a more informed understanding of the discussion at hand.

3.2 Limitations of our model / Ethical Considerations

After addressing the model's objectives in Section 3.1, it is crucial to highlight its limitations and potential ethical concerns. Due to existing technical constraints and ethical factors, our model is presently unsuitable for the public. However, it represents a important step toward developing a widely applicable solution.

Firstly, we should examine the technical constraints. Our recommendation system utilizes diverse machine learning models to evaluate distinct aspects of articles and comments. These models, such as stance detection, continue to undergo extensive research. We will detail the specific models we employ in the following sections. Currently, our system is restricted to topics these models have been trained on, and our future efforts will concentrate on extending recommendations to encompass a broader range of topics.

Secondly, our model does not distinguish between authentic and fake information. At present, it identifies comments that discuss the same subject as the user comment, offering an argument for their position and presenting various perspectives on the topic. As a result, the model might retrieve misleading recommendations that contribute to the spread of false information on topics such as climate change and abortion. This could aid the spread of misinformation, given that [Dixon and Clarke 2013] has shown how presenting debunked information on the same level with verified facts can lead to an impression of false balance. Such practices might amplify readers' doubts about certain topics. Hence, a mechanism is required to clearly indicate that these pieces of information do not hold the same value.

We plan to tackle this issue in future studies using multiple strategies. One proposed method involves continuing to suggest arguments that might contain harmful information, while simultaneously providing fact-checking and marking these statements as *probably false information*, along with sources explaining why they are labeled this way. This strategy to promote diversity in online discussions is the logical extension of our current approach and the work by [Zerback and Schneiders 2024] to help users build a well-founded opinion, as it reveals all perspectives on a subject while highlighting misinformation.

3.3 How does the model work?

3.3.1 Basic concept of the model. As previously described, the model suggests a variety of perspectives from various sources and communities related to a user's comment of interest. An initial solution might be to construct a labeled dataset and train a machine learning model to address this issue. This approach has been demonstrated to be effective and has been employed by numerous researchers, as discussed in Section 2. However, we have chosen an alternative approach for the following reason. It is challenging to find appropriate labels for this problem. We aim to identify relevant recommendations for a given comment that offer divers perspectives on the comments position. However, there may be several potential recommendations that qualify as appropriate. Employing a fixed label or answer set for the model might be to overly restrictive on the range of feasible answers. A model that suggests appropriate comments, yet not the anticipated ones, would be disregarded even if it offers valid responses.

Futhermore, we opted to break down this challenging problem into smaller, more manageable tasks. Upon closer examination, we discern two distinct tasks, which we subsequently divide into even smaller components to discover feasible solutions.

The initial challenge is to identify the comments most pertinent to the context of the specified user remark—these are comments that offer a position along with robust reasoning on the user's topic. However, merely pinpointing the most pertinent comments is insufficient. As pointed out by [Mahajan et al. 2012], focusing solely on locating the most relevant comments may lead to suggesting only repetitive or similar perspectives. Consequently, the second challenge is to ensure that, in addition to identifying the most relevant comments, these comments also need to present diverse perspectives on the topic at hand.

Relevant Comments. The initial question to address is: *How can we determine whether a comment is relevant to a specific user comment?* To tackle this, we have created a **recommendation score**, calculated as the product of the **semantic similarity** between

[1] Implementation(anonymized): https://anonymous.4open.science/r/What-Do-Other-People-Think-About-This—Recommendation-Model-16E4

the user's comment and the potential comment, along with the **relevance score** of the potential comment.

$$recommendation\_score = semantic\_similarity \times relevance\_score$$

The **relevance score** is based on [Steimann et al. 2024] and depicts the argumentative quality of the comment based on its structural features. We will later explain the score in more detail.

Diverse Set of Viewpoints. The second issue we encounter is the array of diverse perspectives. Our task is to identify comments that are relevant to a specific user comment while also presenting varied viewpoints. To achieve this, we form candidate sets from articles that cover the same subject as the one where the user comment was originally posted. These sets are compiled using labels derived from machine learning techniques such as stance detection, sentiment, or emotion classification. For instance, implementing stance detection results in two groups: one comprising comments with a positive stance and another with a negative stance on the article's topic. Subsequently, we rank these groups based on the recommendation score of the comment candidates, select the N most pertinent comments from each group, and merge them into a single set. This unified set thus represents the most relevant and diverse perspectives for the given user comment.

3.3.2 Input and Output of the Recommendation Model. To provide recommendations, the model needs two inputs: the user comment and the keywords associated with the article where the comment has been posted. These keywords encapsulate the article's primary subject, narrowing the focus for recommended comments. This approach greatly reduces the time required to produce recommendations. More information will be presented in the subsequent subsection.

3.3.3 How the model works in detail. In Section 3.3.1, we explained that our model finds relevant comments by assigning a **recommendation score** to each comment candidate which is based on semantic similarity to the user comment and the relevance score of the candidate itself. However, evaluating every possible user-comment candidate pairing in the system is impractical, effectively rendering the system unusable. For this reason, we streamline the selection of comment candidates by first narrowing down the set of candidates by using comments from articles on similar topics before computing the recommendation score. We identify these articles through the keywords of the article where the user's comment is placed. This strategy follows the model suggested by [Steimann et al. 2022].

In order to accomplish this, we create nodes in the database, as seen in Figure 1, for each article and associate them with topic nodes representing the article's keywords. A unique topic node is generated for every keyword. When a new article is introduced, we check if a relevant topic node already exists for each keyword. If it does, the article node is linked to the existing topic node. For retrieval, we search for pertinent topic nodes by using the keywords submitted with the article, incorporating nodes with keywords that are semantically similar[2]. Afterward, we fetch all article nodes from the database that are associated with these topic nodes and further refine this collection by retaining only those articles whose keywords are sufficiently semantically similar to those of the user's article. This approach allows us to effectively narrow down the potential comment pool.

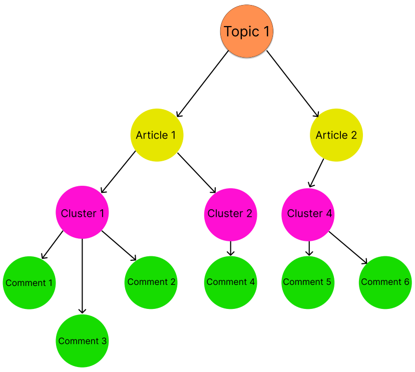
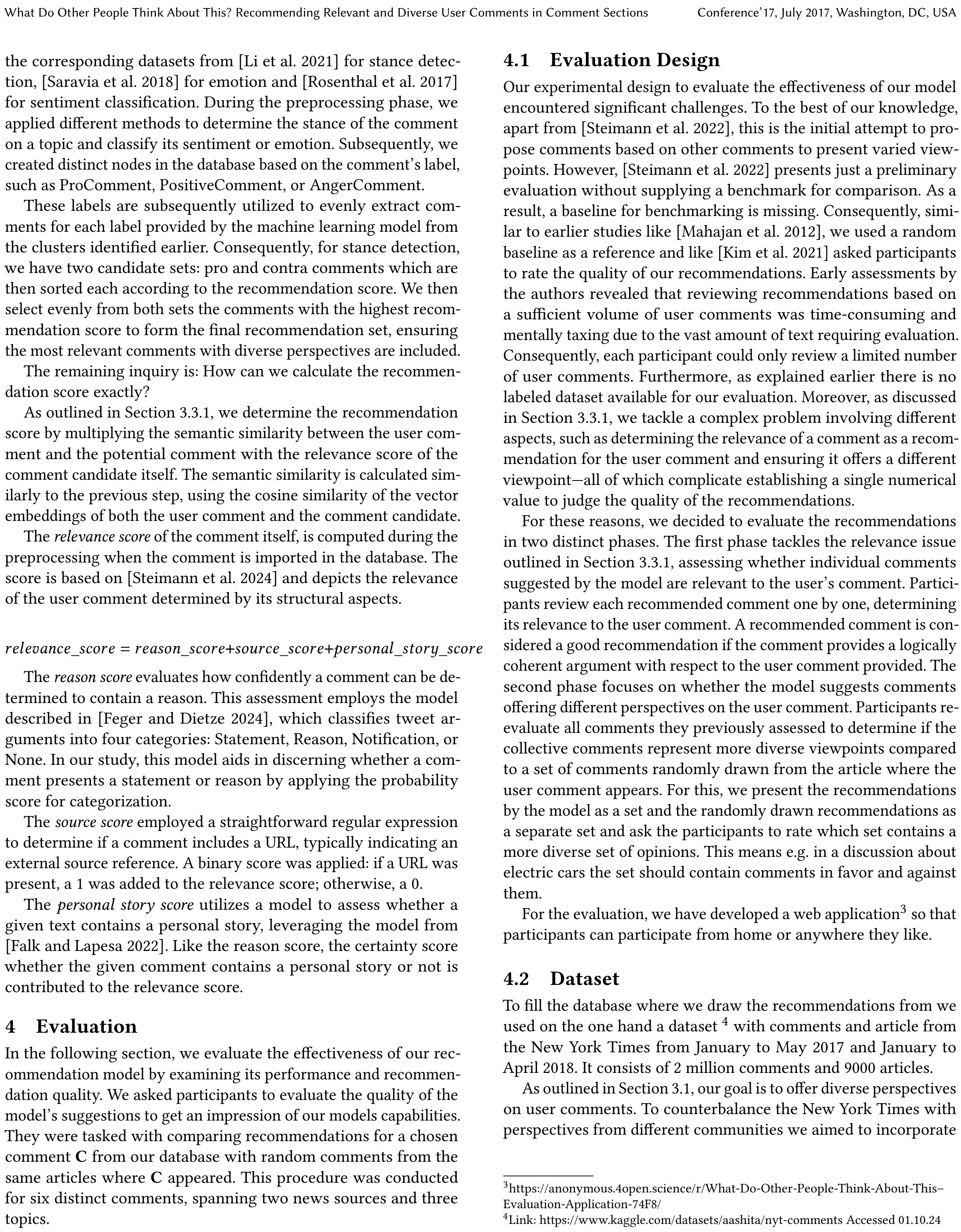


Figure 1: Within the database, the hierarchy begins with the Topic node (orange), linked to the article nodes (yellow), which in turn connect to the cluster nodes (pink), and these are ultimately linked to the comment nodes (green).

We can further narrow down this pool by implementing extra preprocessing measures. We assume that the focus of a discussion evolve over time, introducing new facets of the subject. Take, for instance, a discourse on solar power; it may initially concentrate on the requisite infrastructure, but eventually broaden to encompass other subjects such as alternative energy sources like wind and nuclear power. Consequently, if a user's comment is centered on infrastructure matters, comments extolling the benefits of nuclear power are deemed less pertinent. To manage this, we establish a linking cluster node to bridge articles and comments, ensuring all related comments from a specific sub-discussion are grouped. The cluster node is characterized by the most defining terms from the comments within the cluster. During preprocessing, we utilize the short-text clustering algorithm from [Yin and Wang 2014] to formulate comment clusters in the import phase of an article. The clusters that possess comments fitting for the candidate pool are identified by evaluating the semantic similarity between the vector embedding of the user comment and that of the cluster representation.

Prior to explaining how we select the candidates from the candidate set using the recommendation score, we address another challenge that our model tackles. As outlined in Section 3.1, our goal is to present a diverse set of relevant comments to the user, offering varied perspectives on the user's input. To achieve this, we experimented with machine learning tasks such as stance detection, sentiment and emotion classification and also points of view based on the news agency where the comment was posted. For stance detection, as well as sentiment and emotion classification, we utilized the model from [Li et al. 2021]. The model was trained with

[2] In this work, we adopt the same methodology as described in [Steimann et al. 2022] for determining semantic similarity throughout all steps. We utilize the all-mpnet-base-v2 model from the sentence-transformers package[Reimers and Gurevych 2019] to embed the texts and calculate the cosine similarity between vectors.



the corresponding datasets from [Li et al. 2021] for stance detection, [Saravia et al. 2018] for emotion and [Rosenthal et al. 2017] for sentiment classification. During the preprocessing phase, we applied different methods to determine the stance of the comment on a topic and classify its sentiment or emotion. Subsequently, we created distinct nodes in the database based on the comment's label, such as ProComment, PositiveComment, or AngerComment.

These labels are subsequently utilized to evenly extract comments for each label provided by the machine learning model from the clusters identified earlier. Consequently, for stance detection, we have two candidate sets: pro and contra comments which are then sorted each according to the recommendation score. We then select evenly from both sets the comments with the highest recommendation score to form the final recommendation set, ensuring the most relevant comments with diverse perspectives are included.

The remaining inquiry is: How can we calculate the recommendation score exactly?

As outlined in Section 3.3.1, we determine the recommendation score by multiplying the semantic similarity between the user comment and the potential comment with the relevance score of the comment candidate itself. The semantic similarity is calculated similarly to the previous step, using the cosine similarity of the vector embeddings of both the user comment and the comment candidate.

The *relevance score* of the comment itself, is computed during the preprocessing when the comment is imported in the database. The score is based on [Steimann et al. 2024] and depicts the relevance of the user comment determined by its structural aspects.

$$relevance\_score = reason\_score + source\_score + personal\_story\_score$$

The *reason score* evaluates how confidently a comment can be determined to contain a reason. This assessment employs the model described in [Feger and Dietze 2024], which classifies tweet arguments into four categories: Statement, Reason, Notification, or None. In our study, this model aids in discerning whether a comment presents a statement or reason by applying the probability score for categorization.

The *source score* employed a straightforward regular expression to determine if a comment includes a URL, typically indicating an external source reference. A binary score was applied: if a URL was present, a 1 was added to the relevance score; otherwise, a 0.

The *personal story score* utilizes a model to assess whether a given text contains a personal story, leveraging the model from [Falk and Lapesa 2022]. Like the reason score, the certainty score whether the given comment contains a personal story or not is contributed to the relevance score.

4 Evaluation

In the following section, we evaluate the effectiveness of our recommendation model by examining its performance and recommendation quality. We asked participants to evaluate the quality of the model's suggestions to get an impression of our models capabilities. They were tasked with comparing recommendations for a chosen comment **C** from our database with random comments from the same articles where **C** appeared. This procedure was conducted for six distinct comments, spanning two news sources and three topics.

4.1 Evaluation Design

Our experimental design to evaluate the effectiveness of our model encountered significant challenges. To the best of our knowledge, apart from [Steimann et al. 2022], this is the initial attempt to propose comments based on other comments to present varied viewpoints. However, [Steimann et al. 2022] presents just a preliminary evaluation without supplying a benchmark for comparison. As a result, a baseline for benchmarking is missing. Consequently, similar to earlier studies like [Mahajan et al. 2012], we used a random baseline as a reference and like [Kim et al. 2021] asked participants to rate the quality of our recommendations. Early assessments by the authors revealed that reviewing recommendations based on a sufficient volume of user comments was time-consuming and mentally taxing due to the vast amount of text requiring evaluation. Consequently, each participant could only review a limited number of user comments. Furthermore, as explained earlier there is no labeled dataset available for our evaluation. Moreover, as discussed in Section 3.3.1, we tackle a complex problem involving different aspects, such as determining the relevance of a comment as a recommendation for the user comment and ensuring it offers a different viewpoint—all of which complicate establishing a single numerical value to judge the quality of the recommendations.

For these reasons, we decided to evaluate the recommendations in two distinct phases. The first phase tackles the relevance issue outlined in Section 3.3.1, assessing whether individual comments suggested by the model are relevant to the user's comment. Participants review each recommended comment one by one, determining its relevance to the user comment. A recommended comment is considered a good recommendation if the comment provides a logically coherent argument with respect to the user comment provided. The second phase focuses on whether the model suggests comments offering different perspectives on the user comment. Participants re-evaluate all comments they previously assessed to determine if the collective comments represent more diverse viewpoints compared to a set of comments randomly drawn from the article where the user comment appears. For this, we present the recommendations by the model as a set and the randomly drawn recommendations as a separate set and ask the participants to rate which set contains a more diverse set of opinions. This means e.g. in a discussion about electric cars the set should contain comments in favor and against them.

For the evaluation, we have developed a web application[3] so that participants can participate from home or anywhere they like.

4.2 Dataset

To fill the database where we draw the recommendations from we used on the one hand a dataset [4] with comments and article from the New York Times from January to May 2017 and January to April 2018. It consists of 2 million comments and 9000 articles.

As outlined in Section 3.1, our goal is to offer diverse perspectives on user comments. To counterbalance the New York Times with perspectives from different communities we aimed to incorporate

[3] https://anonymous.4open.science/r/What-Do-Other-People-Think-About-This–Evaluation-Application-74F8/

[4] Link: https://www.kaggle.com/datasets/aashita/nyt-comments Accessed 01.10.24

comments from different sources, such as FoxNews[5]. However, many outlets, including FoxNews, have disabled their comment sections and shifted discussions to social media like X[6], formerly Twitter [Nelson et al. 2021]. To the best of our knowledge, no dataset exists containing both comments and news articles from FoxNews or a comparable news outlet which are necessary for our model. We also choose not to use comments extracted from social networks because this is a similar but still different medium that results in distinct writing style and commenting behavior [Ben-David and Soffer 2019][7][Rowe 2015]. Consequently, we opted to compile a dataset using comments from Breitbart[8]. It is crucial to note we recognize the need for cautious use of Breitbart due to its frequent dissemination of misinformation and hate speech. As elaborated in Section 3.2, our model is not currently suitable for real-world application. In Section 5, we will discuss ideas for handling false information and hate speech. Currently, our aim is to assess whether our model can offer pertinent comments that engage with the positions expressed in user comments and provide alternative viewpoints.

For this experiment, we extracted 135 articles for the topics of *Donald Trump*, *Abortion*, *Jobs and Labor*, *Presidential Race 2024*, *Climate Change*, and *from around the world* with 159323 comments from Breitbart.

4.3 Experiment

For the evaluation process, we selected six user comments from our database. Participants assessed the recommendations produced by our model for these comments in contrast to randomly drawn comments from the same article where the user comment has been posted. The topics of interest were *Donald Trump, Abortion, and Climate Change* because some machine learning models utilized in this study, such as stance detection, were specifically trained on these areas, as explained in Section 3.2. We balanced the representation of comments from both news organizations on each topic. Clear criteria were established to select appropriate comments, ensuring transparency in the selection process while excluding those unsuitable for recommendation, such as comments that were off-topic, displayed incoherent topic shifts, or lacked a clear argument or stance.

1. The comment should clearly align with one of the issues: *Donald Trump, Abortion, or Climate Change.*
2. It must concentrate on a single topic without shifting focus.
3. The comment must present a clear and easy to understand argument.

4.3.1 Demographics. The evaluation involved 11 participants form the computer science department, including 4 women and 7 men. Two participants needed to be excluded due to incomplete evaluations. The remaining Participants' educational backgrounds comprised 5 PhDs and 4 Masters degree. Age distribution included 1 participant aged 20-29, 7 aged 30-39, and 1 aged 50-59. Furthermore, we should note that our sample of participants is not representative for the population and our sample size is insufficient to assert the universality of our findings. Further studies are needed.

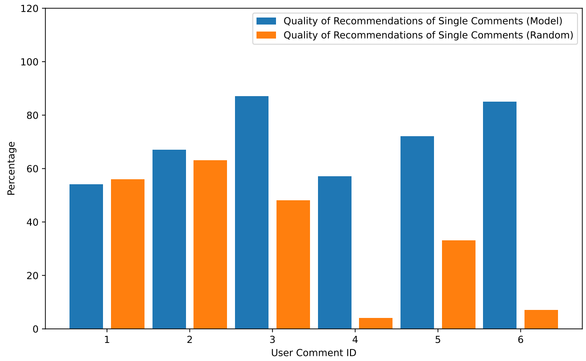


Figure 2: Histogram with percent values how many answers considered the seen comment as a good recommendations for our model vs. the random model.

4.3.2 Results. Our results[9] indicate that our main hypothesis is true and our model performs better than randomly drawn comments for the same article where the user comment has been posted.

This applies to the relevance aspect of the recommended comments as well as the variety of opinions. During the evaluation, participants were tasked with rating the individual recommended comments sequentially to determine if each comment was a suitable suggestion for the corresponding user comment. Figure 2 illustrates the percentage values of answers that assessed individual comments as a *good recommendation.* It shows that our model's recommendations for user comments with IDs 1-2 are considered equally good to randomly selected comments, and for questions with IDs 3-6, they are seen as significantly better. The average percentage for model recommendations is 70.33(±13.79), while random recommendations hold an average of 35.17%(±25.07%).

For the second aspect, as we aim to recommend a diverse set of opinions. Participants were tasked with comparing two sets which consist of the comments they had evaluated in the previous step to determine if one offered a wider range of viewpoints. We present the percentage values on how often our model was preferred over randomly selected comments. Participants could also indicate if both sets were equally good or poor at showcasing differing perspectives. In this case, we counted this as if participants had selected the random set. As seen in Figure 3, the results are similar to the relevance results in Figure 2. We see that our model performs very good for user comments with ID 3-6 like in Figure 2. However, we also see a worse performance for the user comments with ID 1 and 2 in comparison to Figure 2. The randomly drawn recommendations have been perceived as more diverse than our recommendations. The average percentage for the preference of our model over random regarding diversity was 67.5%(±30.02%).

[5] https://www.foxnews.com/

[6] https://x.com/

[7] This study was conducted with Israeli news websites. The results are therefore only transferable to a limited extent. However, it provides evidence in favor of our argument

[8] https://www.breitbart.com/

[9] https://anonymous.4open.science/r/What-do-other-people-think-about-this—Evaluation-Results-05E4/

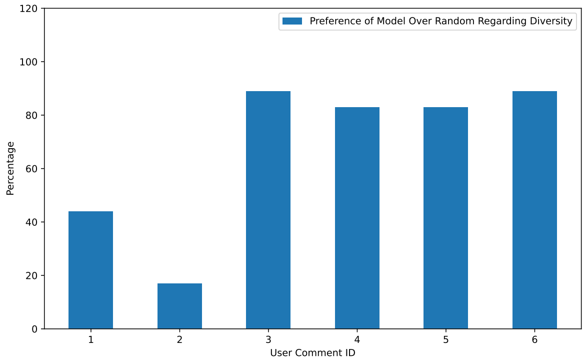


Figure 3: Histogram with percent values for how often our model was preferred over randomly drawn recommendations regarding the diversity of the recommendations.

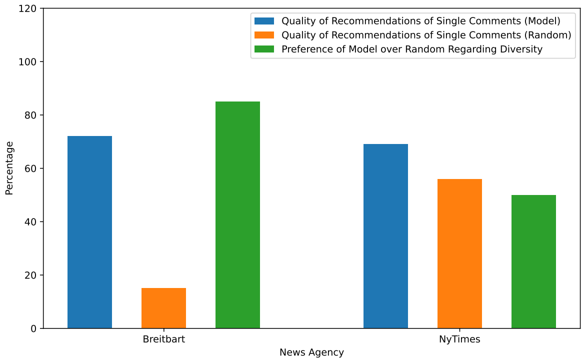


Figure 4: Comparison of good recommendation for singe comments and comparision of viewpoint sets for NyTimes and Breitbart.

As illustrated in Figures 2 and 3, the model performed on par or worse than a randomly selected comment for user comments with IDs 1 and 2, while it outperformed for the other user comments. This observation is intriguing as comments with IDs 1-3 are from NyTimes articles, whereas the others are from Breitbart. Further analysis in Figure 4 shows that for NyTimes comments, the model's performance lagged compared to Breitbart. For comments from the NyTimes, 57% of random suggestions were rated positively by users, with our model slightly outperforming at 69%. In terms of diversity, our model performed equally to the random baseline, being chosen by 52% of the answers. On the other hand, for comments on Breitbart, 72% of our model's recommendations were rated as good compared to only 15% for random suggestions. Furthermore, 85% of responses favored our model over random recommendations regarding recommendation diversity. Consequently, our model shows particularly strong performance for Breitbart, while it only marginally surpasses random suggestions from the NyTimes on average. Nevertheless, as depicted in Figures 2 and 3, the outcomes for NyTimes user comments differ significantly, especially between IDs 1 and 2 compared to ID 3. For ID 3, our model beats random recommendations. This intriguing finding warrants further exploration. We speculate that this discrepancy may stem from the NyTimes's more curated comment sections, which prioritize comment quality and attract an audience with a relatively high percentage of college-educated individuals[Statista 2024]. Thus, the proportion of high-quality contributions is likely greater. This could clarify why our model excels more with Breitbart compared to the NyTimes. Comment sections that are not curated provide an excellent opportunity for our model to suggest valuable contributions with varied viewpoints. Despite this, as demonstrated by user comment with ID 3, our model remains beneficial even in highly curated comment sections belonging to reputable news organizations.

Additionally, we examined whether the perceived relevance of our comments is influenced by the length of the suggested comments. It might be assumed that lengthy comments are inherently seen as more pertinent compared to brief ones. Our findings indicate that only 13% of short comments suggested randomly were deemed relevant. In contrast, 59% of short comments recommended by our model were considered relevant by participants. For lengthy comments, 62% of those randomly suggested were viewed as valuable recommendations. Recommendations from our model were perceived as relevant in 72% of the instances. This suggests that longer comments are generally regarded as more pertinent, as evidenced by the random recommendations. Nevertheless, the results also show the existence of relevant short comments, which our model successfully identifies. In our assessment, comments with fewer than 200 characters were categorized as short.

User Comment ID	Stance	Sentiment	Emotion	News Agencies
1	18.34	18.34	25.53	30.36
2	16.43	17.12	18.92	24.05
3	11.06	11.56	11.78	25.11
4	13.18	13.78	14.21	25.4
5	4.75	4.93	6.21	19.64
6	10.25	11.13	11.98	32.34

Table 1: Overview of performance assessment of the average response times over 10 iterations per user comment and machine learning technique

4.4 Performance Evaluation

In addition, we conducted a performance evaluation of the recommendation model to assess its efficacy in generating recommendations (Table 1). For this purpose, we documented the response times, measured in seconds, necessary to produce recommendations for the user comments used in the participant evaluation. Evidently, the stance detection model yielded the best performance with an average response time of 12.34 seconds, while the sentiment- and emotion-based methods followed with 12.81 and 14.77 seconds, respectively. The news-agency based approach performed slightly worse, but still very well with 26.15 seconds.

4.5 Discussion

As far as we are aware, our model is the first recommendation method for user comments specifically designed to offer varied perspectives on specific user comments besides [Steimann et al. 2022]. We aim that our model serves as a complementary addition to the current research detailed in Section 2, potentially aiding in fostering public discourse and helping users to form a well-founded opinion.

Our results indicate that our recommendation model outperforms randomly drawn comments from the same discussion where the user comment has been posted. We are able to provide relevant and diverse perspectives for the given user comments.

The findings imply that our model could evolve into one suitable for real-world discussions. Nevertheless at present, its applicability is limited to certain topics, and the quality of the recommendation is inconsistent. For instance, the model handles comments from Breitbart more effectively than those from the NyTimes. Further investigation is needed to comprehend the reasons for these disparities and to be able to provide recommendaitons for universal topics.

Additionally, different problems need to be solved, such as the possible promotion of false information and hateful ideas. We will explain possible solutions for these in Section 5.

5 Conclusion and Future Work

In this paper, we proposed a novel recommendation approach for user comment recommendation in the context of news article. Previous work in this field focused on finding the most relevant comment in a discussion or provided tools and visualizations for an exporative overview about different aspects of a discussion.

We want to provide a complementary take on this subject by shifting the focus to the individual comments and ask the question *What do other people think about this comment?*

We introduced our recommendation model, designed to suggest comments from various articles and communities that align with the user's current interests. The purpose of this model is to help users develop comprehensive opinions by providing exposure to a range of perspectives they may otherwise overlook.

Our evaluation showed that our model outperforms the base model, which is a random drawn comment from the same discussion where the comment has been published the user is currently interested in.

Although our model has shown promising results, it remains an interim step. Currently, its applicability is limited to a specific range of topics because the machine learning models employed have been trained only on these areas. To achieve broader usability, this limitation must be addressed. Since these machine learning topics are still subjects of ongoing research, future efforts should focus on exploring alternative approaches that offer greater generalization.

An essential concern for future research is that our current model does not distinguish between misinformation and verified facts, which can be particularly harmful in the public debate. We have identified two potential strategies to address this issue. Firstly, false information or hate speech should be identified and excluded during preprocessing to prevent its recommendation. Secondly, rather than removing these comments, they could be flagged with a warning that indicates potential inaccuracy, along with references supporting this assessment. This second approach is particularly intriguing as it educates users about diverse opinions while clarifying discrepancies with e.g. scientific consensus without excluding other peoples view.

References

Deepak Agarwal, Bee-Chung Chen, and Bo Pang. 2011. Personalized recommendation of user comments via factor models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*. 571–582.

Anat Ben-David and Oren Soffer. 2019. User comments across platforms and journalistic genres. *Information, Communication & Society* 22, 12 (2019), 1810–1829.

Sihao Chen, Daniel Khashabi, Chris Callison-Burch, and Dan Roth. 2019a. PerspectroScope: A window to the world of diverse perspectives. *arXiv preprint arXiv:1906.04761* (2019).

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019b. Seeing things from a different angle: Discovering diverse perspectives about claims. *arXiv preprint arXiv:1906.03538* (2019).

Graham N. Dixon and Christopher E. Clarke. 2013. Heightening Uncertainty Around Certain Science: Media Coverage, False Balance, and the Autism-Vaccine Controversy. *Science Communication* 35, 3 (2013), 358–382. https://doi.org/10.1177/1075547012458290 arXiv:https://doi.org/10.1177/1075547012458290

Neele Falk and Gabriella Lapesa. 2022. Reports of personal experiences and stories in argumentation: datasets and analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 5530–5553. https://doi.org/10.18653/v1/2022.acl-long.379

Marc Feger and Stefan Dietze. 2024. BERTweet's TACO Fiesta: Contrasting Flavors On The Path Of Inference And Information-Driven Argument Mining On Twitter. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 2256–2266. https://doi.org/10.18653/v1/2024.findings-naacl.146

Seth Flaxman, Sharad Goel, and Justin M Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly* 80, S1 (2016), 298–320.

Glen Joris, Camiel Colruyt, Judith Vermeulen, Stefaan Vercoutere, Frederik De Grove, Kristin Van Damme, Orphée De Clercq, Cynthia Van Hee, Lieven De Marez, Veronique Hoste, et al. 2020. News diversity and recommendation systems: Setting the interdisciplinary scene. *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers 14* (2020), 90–105.

Taewook Kim, Hyunwoo Kim, Juho Kim, and Xiaojuan Ma. 2021. Improving readers' awareness of divergent viewpoints by displaying agendas of comments in online news discussions. In *Companion publication of the 2021 conference on computer supported cooperative work and social computing*. 99–103.

Hayato Kobayashi, Hiroaki Taguchi, Yoshimune Tabuchi, Chahine Koleejan, Ken Kobayashi, Soichiro Fujita, Kazuma Murao, Takeshi Masuyama, Taichi Yatsuka, Manabu Okumura, et al. 2021. A Case Study of In-House Competition for Ranking Constructive Comments in a News Service. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. 24–35.

Varada Kolhatkar and Maite Taboada. 2017. Using new york times picks to identify constructive comments. In *Proceedings of the 2017 EMNLP workshop: Natural language processing meets journalism*. 100–105.

Varada Kolhatkar, Nithum Thain, Jeffrey Sorensen, Lucas Dixon, and Maite Taboada. 2020. Classifying constructive comments. *arXiv preprint arXiv:2004.05476* (2020).

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*. 661–670.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2021. Improving Stance Detection with Multi-Dataset Learning and Knowledge Distillation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6332–6345. https://doi.org/10.18653/v1/2021.emnlp-main.511

Tianrui Liu, Changxin Xu, Yuxin Qiao, Chufeng Jiang, and Weisheng Chen. 2024. News recommendation with attention mechanism. *arXiv preprint arXiv:2402.07422* (2024).

Dhruv Kumar Mahajan, Rajeev Rastogi, Charu Tiwari, and Adway Mitra. 2012. Logucb: an explore-exploit algorithm for comments recommendation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. 6–15.

Ankan Mullick, Sayan Ghosh, Ritam Dutt, Avijit Ghosh, and Abhijnan Chakraborty. 2019. Public sphere 2.0: Targeted commenting in online news media. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne,* *Germany, April 14–18, 2019, Proceedings, Part II 41*. Springer, 180–187.

Maria N Nelson, Thomas B Ksiazek, and Nina Springer. 2021. Killing the comments: Why do news organizations remove user commentary functions? *Journalism and Media* 2, 4 (2021), 572–583.

Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1114–1125.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084

Julian Risch, Victor Künstler, and Ralf Krestel. 2020. HyCoNN: hybrid cooperative neural networks for personalized news discussion recommendation. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 41–48.

Julian Risch, Tim Repke, Lasse Kohlmeyer, and Ralf Krestel. 2021. ComEx: Comment Exploration on Online News Platforms.. In *IUI Workshops*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*. 502–518.

Ian Rowe. 2015. Deliberation 2.0: Comparing the Deliberative Quality of Online News User Comments Across Platforms. *Journal of Broadcasting & Electronic Media* 59, 4 (2015), 539–555. https://doi.org/10.1080/08838151.2015.1093482 arXiv:https://doi.org/10.1080/08838151.2015.1093482

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3687–3697. https://doi.org/10.18653/v1/D18-1404

Erez Shmueli, Amit Kagian, Yehuda Koren, and Ronny Lempel. 2012. Care to comment? Recommendations for commenting on news stories. In *Proceedings of the 21st international conference on World Wide Web*. 429–438.

Statista. 2024. Daily newspaper: The New York Times readers in the United States. Available at https://www.statista.com/study/91058/daily-newspapers-the-new-york-times-readers-in-the-united-states/ (2024/12/05).

Jan Steimann, Markus Brenneis, and Martin Mauve. 2024. Is This Comment More Relevant? Understanding the Structural Aspects of Relevance in Comment Sections. In *International Conference on Human-Computer Interaction*. Springer, 264–278.

Jan Steimann, Marc Feger, and Martin Mauve. 2022. Inspiring Heterogeneous Perspectives in News Media Comment Sections. In *Human Interface and the Management of Information: Visual and Information Design*, Sakae Yamamoto and Hirohiko Mori (Eds.). Springer International Publishing, Cham, 118–131.

Diego Uribe, Enrique Cuan, and Elisa Urquizo. 2020. Representation Learning for Constructive Comments Classification. In *2020 International Conference on Mechatronics, Electronics and Automotive Engineering (ICMEAE)*. IEEE, 71–75.

Cedric Waterschoot and Antal van den Bosch. 2024. A time-robust group recommender for featured comments on news platforms. *Frontiers in big Data* 7 (2024), 1399739.

Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 233–242.

Thomas Zerback and Pascal Schneiders. 2024. Noticed and appreciated? The role of argument diversity in enhancing news credibility and reader satisfaction. *Journalism* (2024), 14648849231226030.

A Online Resources

We utilized the Writeful tool available on Overleaf to enhance the quality of our text and employed Phind.com, similar to Google, for addressing coding inquiries throughout the creation of our evaluation web application and the implementation of our recommendation model.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Chapter 7

Lessons Learned and Future Work

This chapter provides an overview where we revisit and analyze the key findings of this thesis, demonstrating how they address the research questions posed in Chapter 1. Following this examination, we consider directions for future research endeavors. Initially, we explore potential approaches to tackle pressing issues such as model evaluation, constraints in topic coverage, and suggestions for handling fake news. Subsequently, we deliberate on a broader scope, focusing on other challenges and the strategies that might address them.

7.1 Lessons Learned

As detailed in Chapter 1, this thesis is a research project with a strong focus on practical applications. Our goal is to explore the viability of a recommendation system tailored for the comment sections of news platforms, addressing the main questions of this thesis: *What do others think about this comment?* Given its application-oriented nature, our emphasis lies in developing a model that not only ensures high-quality recommendations but also excels in practical usability and retrieval effectiveness. We aim to establish the groundwork for future research projects by introducing this novel recommendation task and examining the potential models we can create with the existing research landscape. Through this exploration, we aimed to set a benchmark and discern what is currently effective, highlighting areas where further research should be directed.

Initially, we constructed a straightforward recommendation model rooted in semantic similarity, detailed in Chapter 3. Within this chapter, we also formulated a strategy that serves as the cornerstone for our work in Chapter 6 where we presented a recommendation model to select fitting and diverse recommendations. Through this approach, we verified that we are able to efficiently identify suitable recommendations for our task by utilizing semantic similarity and the two-step retrieval technique outlined in the study.

In the subsequent phase, we improved our emphasis on the practical aspect of our research by creating an open-source framework dedicated to the evaluation of prototypes of comment recommendation systems, as detailed in Section 4. This framework enables the rapid development of prototypes suitable for both controlled laboratory environments and practical real-world applications. Through this framework, it becomes feasible to construct future recommendation systems with an emphasis on practical considerations, allowing convenient modifications to test the model in various scenarios or with different e.g. user-interfaces for the recommendation model. The framework provides a complete technical infrastructure for the recommendation model, encompassing a database, and includes additional systems like a Web crawler to gather comment sections from news sites for new data, along with an import script to add data to the database.

In Section 5, we turn our attention to a crucial issue that must be resolved prior to developing our recommendation model: assessing the relevance of user comments. To address this, we decomposed the intricate question of *How relevant is this user comment?* into manageable subproblems. This approach was inspired by existing research on the relevance of user comments in online discussions. As presented in Rowe (2015), the authors discuss several crucial components of relevant user comments and examine these factors to understand the deliberative quality of comments in a news outlet's comment section compared to its Facebook page. By selecting various aspects of this study, we conducted a user-study to determine how the presence or absence of these structural attributes affects the perceived relevance of user comments. Our findings do not reveal a definitive threshold for characteristics that universally determine the relevance of a comment. Instead, the presence or absence of certain traits alters the perceived relevance. This understanding enables us to formulate a recommendation score in chapter 6. By leveraging the features discussed in Section 5, which can be computed during database import, we streamline the retrieval process, thereby reducing computational complexity. However, it is important to recognize that our results mainly give us a preliminary understanding of how users view structural factors regarding the relevance of comments, primarily because the participant sample size was limited.

In conclusion, we developed our recommendation model by leveraging the insights detailed earlier in this section. We expanded on the findings from Chapter 3, employing semantic similarity and a two-tier retrieval process to efficiently find semantically appropriate recommendations. In addition, the insights from Section 5 and Section 4 were employed to compute a score that determines the relevance of user comments independent of their context during preprocessing, while ensuring the recommendations are timely and application-focused. Our findings suggest that by integrating structural elements with various labels from other machine learning tasks like stance detection or sentiment analysis, we successfully delivered recommendations that evaluation participants regarded as both high-quality and diverse. Nonetheless, it is important to acknowledge that our effective recommendations are confined to specific topics, attributed to the reliance on pretrained machine learning models, and our model does not always outperform the random baseline.

Summary of Key Contributions of This Thesis

1. We introduced the novel task aimed at recommending relevant and diverse comments from comment sections for a specific comment that has captured the user's interest.

2. Our research demonstrated that a simple recommendation model, utilizing semantic similarity alongside keyword based filtering, is capable of suggesting appropriate comments in response to a user comment.
3. We can quickly create realistic prototypes of comment recommendation systems that prioritize application-oriented objectives.
4. We gained insights into how structural features affect the perceived importance of user comments, and this understanding assists us in evaluating a user comment's relevance.
5. We successfully demonstrated that employing a model grounded in structural characteristics, alongside well-known machine learning tasks like stance detection, sentiment analysis, and emotion classification, is capable of generating promising initial outcomes.

7.2 Future Challenges And Possible Solutions

In the preceding section, we outlined the insights and accomplishments presented in this thesis. Nevertheless, this marks only the initial phase in creating a universally accessible application intended for public use. As detailed in the paper discussed in Chapter 6, this endeavor represents merely an intermediate stage towards developing a publicly usable recommendation system. Along this path, we must confront various challenges, and we plan to explore potential solutions for these issues in this section. Initially, we will delve into problems which are more important to address which are identified during this research and afterwards discuss long-term objectives and concepts.

Problems and Possible Solutions

Some of the main open challenges of this thesis appeared during the development of the last paper, described in Chapter 6.

The Evaluation It was very difficult to evaluate our recommendation model. As explained in Chapter 6, we tried to solve the main challenge of finding revenant and diverse recommendations by dividing this task into more manageable sub tasks. In the paper, we obtain an insight about the quality of our model by conducting an evaluation with participants who assessed the quality of our recommendations in terms of relevance and diversity. However, this is hardly a sustainable solutions for future developments of our model due to the fact that evaluations of this kind are very time- and work-consuming, especially if we want to find representative results. Therefore, we need other ways to assess the quality of future recommendation models.

A possible solution might involve reevaluating our assumptions regarding dataset usage, taking cues from the methodology outlined in Chen et al. (2019b). In that study, the authors introduce a dataset designed to find perspectives regarding a specific claims. Their dataset includes labels that consist of a collection of accepted perspectives, and they assess their models based on well-known precision and recall metrics. We could adopt a comparable strategy by creating a dataset comprising user comments alongside corresponding recommendations. Moreover, to address the concerns raised in Steimann and Mauve (2025) about omitted recommendation models that propose appropriate comments absent in the expected recommendation set, we could consider these approaches: firstly, assessing the viability of using a semantic similarity score to contrast model-generated recommendations with our reference dataset; alternatively, investigating data augmentation methods to broaden the spectrum of acknowledged recommendations in our dataset using synonyms or paraphrases; or another intriguing idea could be to employ an LLM to evaluate the recommendation models automatically, and if the outcomes are promising, proceed to evaluate the model with actual users. However, the approach of developing a gold-standard dataset for recommendations is only possible if we find a solution for another problem first, the data basis.

Data Basis At present, our database lacks a fair representation of comments and articles from various communities. We have an extensive collection of comments from the NyTimes[1], but only a limited amount of data from sources like Breitbart[2]. Balancing the NyTimes comments is crucial. Thus, we need either an alternative data source or to acquire more Breitbart comments. As noted in Chapter 6, caution is advised when using Breitbart due to minimal moderation and frequent presentation of false information or hate speech. Complicating this is the fact that many news platforms have disabled comment sections, opting for X(formally Twitter)[3] (Nelson et al., 2021). Consequently, we must look for existing datasets or consider exploring non-English languages. In Germany, numerous news platforms host vibrant comment sections across the political spectrum. Considering the broader data available, examining German comment sections might be a viable strategy.

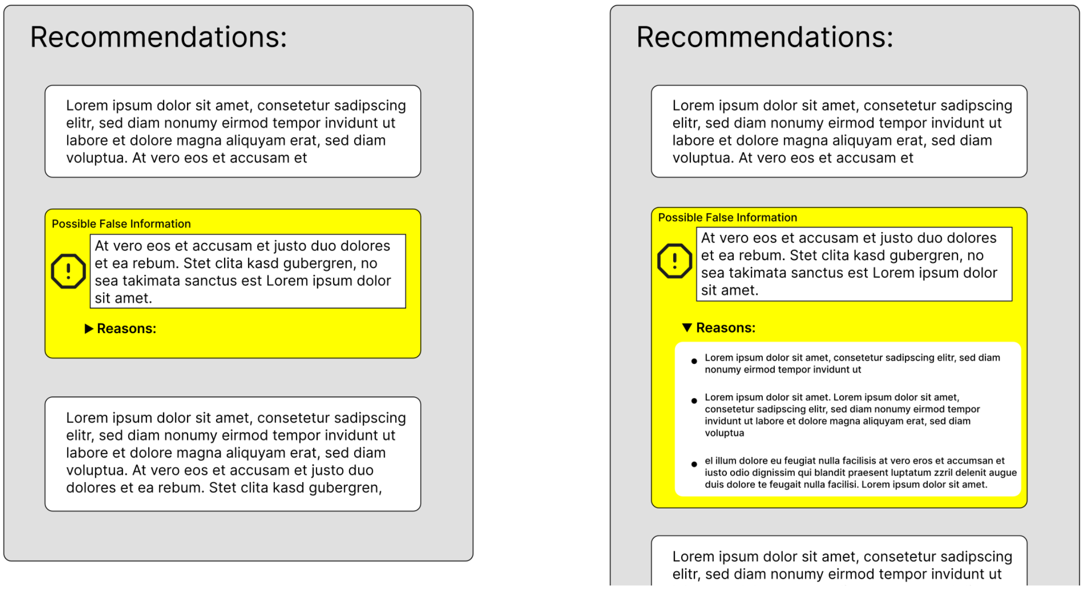
Topic Limitations At present, our recommendation model, presented in Chapter 6, is limited to three topics: *Donald Trump, Abortion, and Climate Change*, because it relies on pre-trained machine learning models. This limitation poses a significant challenge that needs addressing. There are multiple potential solutions for this issue. Primarily, we could concentrate on the machine learning models employed and explore ways to generalize them. This could involve training them with additional datasets or examining different research methodologies that tackle these machine learning tasks. If achieving this is not possible in the foreseeable future, an alternative approach would be to adapt our recommendation model by exploring other methodologies to enhance diversity.

Recommendation of False Information As demonstrated by Dixon and Clarke (2013), the dissemination of false information arises when misinformation is equally presented to verified facts. Although their analysis centeres on articles, it is plausible to extend these findings to comments to some extent. Consequently, as discussed in Chapter 6 regarding the use of

[1] https://www.nytimes.com/

[2] https://www.breitbart.com/

[3] https://twitter.com/



Recommendations:

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et

Possible False Information

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

▶ Reasons:

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren,

Recommendations:

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et

Possible False Information

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

▼ Reasons:

- Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut
- Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua
- el illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut

Figure 7.1: An illustration showcasing a user interface designed for highlighting misinformation along with the related justifications

Breitbart as a data source, this presents a challenge for us. We must develop strategies to manage false information and harmful comments. In Steimann and Mauve (2025), we have already outlined some concepts that we aim to elaborate on here. Primarily, we require an automated system for assessing comments to identify misleading content and hateful comments. Subsequently, we need strategies for handling these. Initially, it seems logical to omit such content during database import; however, automatically filtering user opinions might hinder user trust. This is especially true in communities where exposing individuals to a broader spectrum of perspectives could be most beneficial.

Research by Zerback and Schneiders (2024) indicates that displaying diverse arguments can enhance the credibility of online news by providing readers with a comprehensive array of reliable information for informed opinion formation. While these results pertain to news articles, it is reasonable to suggest the potential applicability of this method to our approach, warranting further investigation. Retaining contentious material while incorporating automated fact-checking and reasoning techniques, as outlined in Guo et al. (2022), allows us to highlight and explain the rationale behind labeling certain statements as misinformation, thereby fostering user trust and dispelling false information. An example user-interface for this approach is illustrated in Figure 7.1. Nevertheless, this issue transcends computer science; identifying and justifying content is merely one aspect. It is equally crucial to discern which method is most effective in elevating credibility and assisting users in shaping their opinions. Thus, collaboration with the social sciences department to explore this further would be invaluable.

User-Interface Besides the essential and imminent challenges previously discussed that must be addressed to develop a universally applicable recommendation system, there are additional compelling aspects that merit further investigation, such as the user interface. A notable area of inquiry involves determining the optimal method of presenting the recommended comments. Thanks to the REST API of our comment recommendation system detailed in Section 4, we possess the flexibility to utilize various user interfaces. In Steimann and Mauve (2024), we proposed the concept of a browser extension to display the recommended comments. A limitation of this interface, however, is that presenting recommendations in a separate window may cause them to appear disconnected from the comments they pertain to. Accordingly, we could expand on this concept by e.g. configuring the browser extension to embed the recommended comments directly into the HTML code, thereby rendering them within the article's comment section. Additional user-interface components can be envisioned, such as offering more detailed information regarding various user comments on a distinct website, along with information about the discussions in which these comments were posted, among other possibilities. Conducting a user-study in this context would be valuable to observe the impact of different user interfaces on the users.

Ration of Supporting and Opposing Views In future research, a noteworthy subject to explore is the balance between supporting and opposing perspectives. Initially, one might assume an equal distribution of supporting and opposing views is ideal. Nevertheless, this assumption warrants scrutiny: If it doesn't mirror the actual distribution, does it remain valid? Moreover, should misinformation receive equal treatment as factual content if this represents the actual ratio? These are pivotal questions that require thorough examination. Additionally, much like tackling the challenge of misinformation, collaborating with experts in social sciences could be beneficial.

Summary of Key Challenges And Future Projects

1. **Evaluation:** It is crucial to identify a method for thoroughly assessing our recommendation model prototypes without the dependence on participant-based evaluations.
2. **Data Basis:** Enhancing our data basis is essential to ensure they adequately reflect a diverse range of perspectives and topics, thereby optimizing the training and evaluation phases of our models.
3. **Topic Limitations:** Addressing the current technical constraints of our model is necessary. Primarily, this involves overcoming the restrictions to a limited number of topics.
4. **False Information and Hate Speech:** We must develop effective strategies for managing false information and hate speech, aiming to maintain a balanced representation of opinions while preventing the equal promotion of misinformation alongside factual content.

5. **User-Interfaces:** Exploring the impact of various user-interface designs on how users perceive the quality of recommendations holds considerable interest and warrants further investigation.
6. **Opinion Ratio:** Examining the best proportion for showcasing various perspectives on a specific comment is an engaging research question that warrants further exploration.

Chapter 8

Conclusion

In this thesis, we explored the problem of discovering appropriate comment recommendations for a user's current comment of interest. To address this, we examined potential ways to develop models that can be formulated given the existing research landscape, emphasizing the application-focused aspect described in Chapter 1. In pursuit of this, we tackled several challenges, such as efficiently identifying thematically relevant recommendations. We also created an open-source software framework designed to facilitate the swift development of recommendation system prototypes, further considering the application-oriented focus. Furthermore, a user study was carried out to evaluate the impact of structural elements on the perceived relevance of comments in comment sections. Ultimately, a recommendation model was developed, synthesizing and expanding upon the findings of prior work.

During the work for this thesis three paper were accepted and published in peer-reviewed conference proceedings: Steimann et al., 2024, 2022; Steimann and Mauve, 2024. Furthermore, the latest paper presenting the recommendation model (Steimann and Mauve, 2025) is currently submitted to a conference.

As described in Chapter 1, a key achievement of this thesis was to lay the foundation along with the identification of challenges and future directions. In this thesis, we address multiple challenges in creating a recommendation model. The fruition of these endeavors is presented in Chapter 6, where we introduce an initial model that delivers relevant and varied insights on user comments.

8.1 Future Work

As discussed in Section 7.2, we have pinpointed several principal challenges that must be tackled in forthcoming research to enable our system's widespread usability. Primarily, it is crucial to explore improved and more efficient methodologies for assessing new iterations of our model. Additionally, we must confront the inherent constraints of the model, such as the topic limitation. Beyond these urgent issues, there are numerous intriguing avenues for further exploration. How should we tackle false information? What is the optimal ratio for presenting opposing viewpoints? Which user interface maximizes effectiveness in displaying recommendations? These queries, among others, warrant investigation in future work.

8.2 Closing Thoughts

In recent years, our focus has been on developing a system to enhance online discussions by helping users form well-rounded opinions and also to foster a better mutual understanding by providing access to diverse perspectives and personal experiences. During our research, we have noted that current online debates frequently feature misleading information, personal attacks, and in the case of Breitbart, blatant hostility toward political opponents. While we recognize that our research cannot address or solve every problem encountered in online discussions, we aspire to support users. Our study aims to enhance empathy in users or offer alternative perspectives by illuminating the viewpoints of others on the matter.

The positive feedback received from visitors during the *Night of Science*[1], where an early version of our recommendation system was presented, confirms strong user interest in this tool.

We fully recognize that in addition to the scientific challenges outlined in Section 7.2, there are other hurdles to consider. For example, to offer this system to the public, continuous access to leading news platforms and their comment areas is essential, allowing us to suggest topics of current interest.

[1] https://www.nachtderwissenschaft-duesseldorf.de/en/night-of-science

Appendix

AI Tools Used In This Thesis

During this thesis, different AI tools were employed to assist our work. Initially, we utilized Writefull[2] on Overleaf[3] to enhance phrasing and to identify grammar and spelling errors during writing of our paper and this thesis. Furthermore, Phind[4] served as an AI query tool for specific questions regarding software libraries and frameworks, such as *How Do I use multiple match clauses in Neo4j Query?*. Lastly, DeepL[5] was applied for translating phrases between English and German.

[2] https://www.writefull.com/

[3] https://de.overleaf.com/

[4] https://www.phind.com/

[5] https://www.deepl.com/de/translator

Bibliography

Agarwal, Deepak, Bee-Chung Chen, and Bo Pang (2011). “Personalized recommendation of user comments via factor models”. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 571–582 (cit. on pp. 1, 8).

Brenneis, Markus and Martin Mauve (Nov. 2020). “Do I Argue Like Them? A Human Baseline for Comparing Attitudes in Argumentations”. In: *Proceedings of the Workshop on Advances In Argumentation In Artificial Intelligence 2020* (Nov. 25, 2020). Ed. by Bettina Fazzinga, Filippo Furfaro, and Francesco Parisi. CEUR Workshop Proceedings 2777. Aachen, pp. 1–15. URL: `http://ceur-ws.org/Vol-2777/paper21.pdf` (cit. on p. 48).

Chen, Sihao, Daniel Khashabi, Chris Callison-Burch, and Dan Roth (2019a). “PerspectroScope: A window to the world of diverse perspectives”. In: *arXiv preprint arXiv:1906.04761* (cit. on pp. 10, 11).

Chen, Sihao, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth (2019b). “Seeing things from a different angle: Discovering diverse perspectives about claims”. In: *arXiv preprint arXiv:1906.03538* (cit. on pp. 10, 11, 79).

Diakopoulos, Nicholas and Mor Naaman (2011). “Towards quality discourse in online news comments”. In: *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*. CSCW ’11. Hangzhou, China: Association for Computing Machinery, pp. 133–142. ISBN: 9781450305563. DOI: `10.1145/1958824.1958844`. URL: `https://doi.org/10.1145/1958824.1958844` (cit. on p. 11).

Dixon, Graham N. and Christopher E. Clarke (2013). “Heightening Uncertainty Around Certain Science: Media Coverage, False Balance, and the Autism-Vaccine Controversy”. In: *Science Communication* 35.3, pp. 358–382. DOI: `10.1177/1075547012458290`. eprint: `https://doi.org/10.1177/1075547012458290`. URL: `https://doi.org/10.1177/1075547012458290` (cit. on p. 80).

Flaxman, Seth, Sharad Goel, and Justin M Rao (2016). “Filter bubbles, echo chambers, and online news consumption”. In: *Public opinion quarterly* 80.S1, pp. 298–320 (cit. on pp. 1, 5).

Guo, Zhijiang, Michael Schlichtkrull, and Andreas Vlachos (2022). “A survey on automated fact-checking”. In: *Transactions of the Association for Computational Linguistics* 10, pp. 178–206 (cit. on p. 81).

Joris, Glen, Camiel Colruyt, Judith Vermeulen, Stefaan Vercoutere, Frederik De Grove, Kristin Van Damme, Orphée De Clercq, Cynthia Van Hee, Lieven De Marez, Veronique Hoste, et al. (2020). “News diversity and recommendation systems: Setting the interdisciplinary scene”. In: *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP*

WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers 14, pp. 90–105 (cit. on p. 9).

Kim, Taewook, Hyunwoo Kim, Juho Kim, and Xiaojuan Ma (2021). "Improving readers' awareness of divergent viewpoints by displaying agendas of comments in online news discussions". In: *Companion publication of the 2021 conference on computer supported cooperative work and social computing*, pp. 99–103 (cit. on pp. 10, 11).

Kobayashi, Hayato, Hiroaki Taguchi, Yoshimune Tabuchi, Chahine Koleejan, Ken Kobayashi, Soichiro Fujita, Kazuma Murao, Takeshi Masuyama, Taichi Yatsuka, Manabu Okumura, et al. (2021). "A Case Study of In-House Competition for Ranking Constructive Comments in a News Service". In: *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pp. 24–35 (cit. on pp. 1, 9).

Kolhatkar, Varada and Maite Taboada (2017). "Using new york times picks to identify constructive comments". In: *Proceedings of the 2017 EMNLP workshop: Natural language processing meets journalism*, pp. 100–105 (cit. on p. 9).

Kolhatkar, Varada, Nithum Thain, Jeffrey Sorensen, Lucas Dixon, and Maite Taboada (2020). "Classifying constructive comments". In: *arXiv preprint arXiv:2004.05476* (cit. on p. 9).

Li, Qing, Jia Wang, Yuanzhu Peter Chen, and Zhangxi Lin (2010). "User comments for news recommendation in forum-based social media". In: *Information Sciences* 180.24, pp. 4929–4939 (cit. on p. 8).

Liu, Tianrui, Changxin Xu, Yuxin Qiao, Chufeng Jiang, and Weisheng Chen (2024). "News recommendation with attention mechanism". In: *arXiv preprint arXiv:2402.07422* (cit. on p. 8).

Mahajan, Dhruv Kumar, Rajeev Rastogi, Charu Tiwari, and Adway Mitra (2012). "Logucb: an explore-exploit algorithm for comments recommendation". In: *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 6–15 (cit. on pp. 1, 9).

Mullick, Ankan, Sayan Ghosh, Ritam Dutt, Avijit Ghosh, and Abhijnan Chakraborty (2019). "Public sphere 2.0: Targeted commenting in online news media". In: *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II 41*. Springer, pp. 180–187 (cit. on pp. 9, 10).

Nelson, Maria N., Thomas B. Ksiazek, and Nina Springer (2021). "Killing the Comments: Why Do News Organizations Remove User Commentary Functions?" In: *Journalism and Media* 2.4, pp. 572–583. ISSN: 2673-5172. DOI: `10.3390/journalmedia2040034`. URL: `https://www.mdpi.com/2673-5172/2/4/34` (cit. on p. 80).

Park, Deokgun, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist (2016). "Supporting comment moderators in identifying high quality online news comments". In: *Proceedings* *of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 1114–1125 (cit. on pp. 8, 9).

Risch, Julian, Victor Künstler, and Ralf Krestel (2020). "HyCoNN: hybrid cooperative neural networks for personalized news discussion recommendation". In: *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, pp. 41–48 (cit. on pp. 8, 10).

Risch, Julian, Tim Repke, Lasse Kohlmeyer, and Ralf Krestel (2021). "ComEx: Comment Exploration on Online News Platforms." In: *IUI Workshops* (cit. on pp. 10, 11).

Rowe, Ian (2015). "Deliberation 2.0: Comparing the Deliberative Quality of Online News User Comments Across Platforms". In: *Journal of Broadcasting & Electronic Media* 59.4, pp. 539–555. DOI: 10.1080/08838151.2015.1093482. eprint: https://doi.org/10.1080/08838151.2015.1093482. URL: https://doi.org/10.1080/08838151.2015.1093482 (cit. on pp. 6, 48, 78).

Shmueli, Erez, Amit Kagian, Yehuda Koren, and Ronny Lempel (2012). "Care to comment? Recommendations for commenting on news stories". In: *Proceedings of the 21st international conference on World Wide Web*, pp. 429–438 (cit. on p. 8).

Sina Blassnig Sven Engesser, Nicole Ernst and Frank Esser (2019). "Hitting a Nerve: Populist News Articles Lead to More Frequent and More Populist Reader Comments". In: *Political Communication* 36.4, pp. 629–651. DOI: 10.1080/10584609.2019.1637980. eprint: https://doi.org/10.1080/10584609.2019.1637980. URL: https://doi.org/10.1080/10584609.2019.1637980 (cit. on p. 7).

Steimann, Jan, Markus Brenneis, and Martin Mauve (2024). "Is This Comment More Relevant? Understanding the Structural Aspects of Relevance in Comment Sections". In: *Human Interface and the Management of Information*. Ed. by Hirohiko Mori and Yumi Asahi. First published in Human Interface and the Management of Information, Page 264–278, 2024 and reproduced with permission from Springer Nature. Cham: Springer Nature Switzerland, pp. 264–278. ISBN: 978-3-031-60107-1 (cit. on pp. 3, 47, 66, 85, 95).

Steimann, Jan, Marc Feger, and Martin Mauve (2022). "Inspiring Heterogeneous Perspectives in News Media Comment Sections". In: *Human Interface and the Management of Information: Visual and Information Design*. Ed. by Sakae Yamamoto and Hirohiko Mori. First published in Human Interface and the Management of Information: Visual and Information Design, Page 118–131, 2022 and reproduced with permission from Springer Nature. Cham: Springer International Publishing, pp. 118–131. ISBN: 978-3-031-06424-1 (cit. on pp. 13, 65, 66, 85, 95).

Steimann, Jan and Martin Mauve (2024). "Developing Custom-Made Comment-Recommendation Prototypes with a Modular Design Framework". In: *Social Computing and Social Media*. Ed. by Adela Coman and Simona Vasilache. First published in Social Computing and Social Media, Page 97–112, 2024 and reproduced with permission from Springer Nature. Cham:

Springer Nature Switzerland, pp. 97–112. ISBN: 978-3-031-61281-7 (cit. on pp. 3, 29, 82, 85, 95).

Steimann, Jan and Martin Mauve (2025). *What Do Other People Think About This? Recommendation Relevant and Divers User Comments in Comment Section.* Manuscript submitted for publication (cit. on pp. 65, 80, 81, 85, 95).

Uribe, Diego, Enrique Cuan, and Elisa Urquizo (2020). “Representation Learning for Constructive Comments Classification”. In: *2020 International Conference on Mechatronics, Electronics and Automotive Engineering (ICMEAE).* IEEE, pp. 71–75 (cit. on p. 9).

Waterschoot, Cedric and Antal van den Bosch (2024). “A time-robust group recommender for featured comments on news platforms”. In: *Frontiers in big Data* 7, p. 1399739 (cit. on p. 8).

Zerback, Thomas and Pascal Schneiders (2024). “Noticed and appreciated? The role of argument diversity in enhancing news credibility and reader satisfaction”. In: *Journalism* 25.12, pp. 2523–2542. DOI: 10.1177/14648849231226030. eprint: https://doi.org/10.1177/14648849231226030. URL: https://doi.org/10.1177/14648849231226030 (cit. on pp. 9, 81).

Zhou, Mingnan, Ruisheng Shi, Zhaozhen Xu, Yuan He, Yiyi Zhou, and Lina Lan (2015a). “Design of Personalized News Comments Recommendation System”. In: *Data Science.* Ed. by Chengqi Zhang, Wei Huang, Yong Shi, Philip S. Yu, Yangyong Zhu, Yingjie Tian, Peng Zhang, and Jing He. Cham: Springer International Publishing, pp. 1–5. ISBN: 978-3-319-24474-7 (cit. on pp. 1, 8).

– (2015b). “Design of personalized news comments recommendation system”. In: *Data Science: Second International Conference, ICDS 2015, Sydney, Australia, August 8-9, 2015, Proceedings 2.* Springer, pp. 1–5 (cit. on p. 29).

Personal Publications

Peer-Reviewed Conference Papers

Jan Steimann, Marc Feger, and Martin Mauve (2022). “Inspiring Heterogeneous Perspectives in News Media Comment Sections”. In: *Human Interface and the Management of Information: Visual and Information Design.* Ed. by Sakae Yamamoto and Hirohiko Mori. First published in Human Interface and the Management of Information: Visual and Information Design, Page 118–131, 2022 and reproduced with permission from Springer Nature. Cham: Springer International Publishing, pp. 118–131. ISBN: 978-3-031-06424-1

Jan Steimann and Martin Mauve (2024). “Developing Custom-Made Comment-Recommendation Prototypes with a Modular Design Framework”. In: *Social Computing and Social Media.* Ed. by Adela Coman and Simona Vasilache. First published in Social Computing and Social Media, Page 97–112, 2024 and reproduced with permission from Springer Nature. Cham: Springer Nature Switzerland, pp. 97–112. ISBN: 978-3-031-61281-7

Jan Steimann, Markus Brenneis, and Martin Mauve (2024). “Is This Comment More Relevant? Understanding the Structural Aspects of Relevance in Comment Sections”. In: *Human Interface and the Management of Information.* Ed. by Hirohiko Mori and Yumi Asahi. First published in Human Interface and the Management of Information, Page 264–278, 2024 and reproduced with permission from Springer Nature. Cham: Springer Nature Switzerland, pp. 264–278. ISBN: 978-3-031-60107-1

Manuscript Submitted for Publication

Jan Steimann and Martin Mauve (2025). *What Do Other People Think About This? Recommendation Relevant and Divers User Comments in Comment Section.* Manuscript submitted for publication

List of Figures

List of Tables

Eidesstattliche Erklärung
laut §5 der Promotionsordnung vom 15.06.2018

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.

Ort, Datum

Jan Lukas Steimann